The Role of Information in Real Estate Markets

DISSERTATION

zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

Eingereicht an der Fakultät für Wirtschaftswissenschaften

der Universität Regensburg

vorgelegt von

Katrin Kandlbinder (M.Sc.)

Tag der Disputation: 12. Juli 2018

Berichterstatter: Prof. Dr. Wolfgang Schäfers

Prof. Dr. Tobias Just

The Role of Information in Real Estate Markets

Katrin Kandlbinder

Table of Contents

Exhibits

1 Introduction

1.1 General Motivation and Theoretical Foundation

The rise of the internet over the last two decades has radically changed the production, availability, distribution, and consumption of information. This information revolution has facilitated not only the access to information by reducing search costs, but also increased the extent of available information (Vlastakis and Markellos, 2012). Due to the enormous growth of web-enabled mobile devices, the internet has become the leading information channel. Everyone is potentially online everywhere at any time. Websites, search engines, and social media act as comprehensive information-exchange tools. Nowadays, a few clicks are enough to gain access to any news source in the world, which is updated by the minute. Along with the enormous increase in information availability, the technological infrastructure has improved as well, providing new tools for analyzing these massive datasets.

One area in which information plays a particularly important role is that of finance. Here, information is regarded as a valuable and highly sought asset, because it is well established that better informed investors are able to earn higher returns. In contrast to efficient market theorists like Fama (1970), it has been recognized that information is imperfect and obtaining it can be costly. Hence, prices do not fully reflect all available information and there are substantial information asymmetries among market participants. Moreover, due to imperfect information about fundamentals, most researchers argue that prices are also influenced substantially by investor sentiment. According to Baker and Wurgler (2007), "the question is no longer, as it was few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and quantify its effects."

With the ongoing creation of a new, quantifiable world, traditional approaches require modification so as to conform to this new environment. In other words, together with the increased availability of unprecedentedly large data volumes, new methods and research fields have evolved, with the potential to make qualitative factors like information and sentiment more tangible and measurable. One of the key features is a rapidly growing amount of textual sources – be they digital-born, such as Tweets, or digitized, such as historical newspapers.

For example, stock message boards (Antweiler and Frank, 2004), digitized US Congressional records (Gentzkow and Shapiro, 2010), or Twitter messages (Lüdering and Tillmann, 2016) are now readily available. They are used in combination with text mining methods like automated content analysis (such as topic modeling), in order to investigate the impact of qualitative information on market developments. Additionally, new sentiment sources like Google search volumes, texts on social media or product reviews, have emerged as well. It has been argued that search engine queries for specific keywords are linked to a wide range of real-world events, and that they are valuable for predicting, for example, unemployment rates (Askitas and Zimmermann, 2009), trading volumes and stock market prices (Da *et al.*, 2014), or even residential or commercial real estate prices (Hohenstatt *et al.*, 2011; Dietzel *et al.*, 2014). Moreover, different text mining methods, such as the dictionary-based approach, support vector machine, or neuronal networks, can be applied to analyze any kind of textual data sources to extract not only essential information, but also sentiment.

These profound changes in information have also affected the real estate industry. The launch of various home search websites, investment platforms, and detailed information about real estate-related investment products, have substantially increased market transparency. However, in the field of real estate, these new research opportunities have not been used with the same intensity as in the finance sector. Yet, information and sentiment are particularly relevant in highly segmented and informationally inefficient markets, such as real estate (Clayton *et al.*, 2009). Asset heterogeneity, infrequent trading, and high transaction costs in the direct real estate market, for example, lead to substantial information asymmetries among market participants, which result in higher price dispersion. As real estate markets are even more prone to information deficits than financial markets, they offer great potential for research which incorporates information and sentiment as measurable factors (Mori, 2015).

Hence, the overall research aim of this dissertation is to investigate three different informational aspects, namely information demand, information availability, and information supply, and their impact and predictive abilities with respect to both direct and indirect real estate markets. Nowadays, people rely on search engines to locate appropriate information in the web. Hence, the first paper concentrates on information demand by using intraday Google search volumes as a proxy for sentiment. The aim is to examine whether there is a relationship between search queries provided by Google Trends and future MSCI Real Estate Investment Trust (REIT) price movements. The second study deals with potential pricing effects caused by increased information availability in US housing markets. It is investigated whether out-of-town buyers pay significantly more for comparable housing, due to the fact that they might be informationally disadvantaged. In order to replicate a sufficient time gap with an enormous improvement in information availability, the findings of two years (2005 and 2015) are compared. Finally, the third paper focuses on the supply side of information by using German real estate newspaper articles for sentiment extraction. As there is no German discipline-specific word list, the first objectively validated *German Real Estate Sentiment Dictionary* was developed, which enables a dictionary-based analysis of German real estate-related text corpora. The resulting sentiment measures are then tested with regard to their predictive abilities for real estate housing price movements.

1.2 Research Questions

This section serves as a basic framework and outlines the research questions addressed in the three papers comprising this dissertation.

Paper 1 | Intraday Online Information Demand and its Relationship with REIT Prices

- Is it possible to predict intraday REIT price movements by using Google search volumes as a sentiment measure?
- Can trading strategies based on changes in Google search volumes outperform a simple buy-and-hold strategy?
- During which market phases (falling, stagnant, or rising) does the Google trading strategy yield higher returns?
- Before which trading signal (buy or sell) does information procurement have the best prediction ability for REITs?
- Are changes in REIT price movements caused by changes in search volumes or vice versa?
- For which asset class does the Google trading strategy perform better – REITs or DJIA stocks?

Paper 2 | Leveling the Playing Field:
Out-of-Town Buyer Premiums in US Housing Markets Over Time

- Were there any changes in the information level and information availability over the last decade due to the internet revolution?
- Do out-of-town buyers pay a premium for real estate compared to their local counterparts?
- If so, is that premium caused by physical distance, anchoring or different personal income levels?
- How do prices react theoretically to changes in search costs and biased beliefs (anchoring)?
- Did the premium caused by distance (search costs) decrease from 2005 to 2015?
- Does the out-of-town buyer premium still exist for propensity score matched samples, which correspond each other regarding housing characteristics?
- Are the findings replicable for other US counties compared to Miami Dade County?

Paper 3 | Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany

- Which German words contain sentiment relating to real estate?
- Do sentiment measures based on the self-created dictionary have predictive power on German residential market returns?
- Is there a causality flow from changes in sentiment to changes in real estate returns or vice versa?
- How crucial is the amount of sentiment words regarding the construction of the dictionary?
- Is the analysis of the headline alone already enough to capture sentiment or does the inclusion of further text lead to better sentiment predictability?
- Do discipline-specific dictionaries produce sentiment measures which more accurately predict subsequent market returns than general ones?
- Are sentiment-augmented VAR models superior to non-sentiment models in terms of forecasting accuracy?

1.3 Course of Analysis

The following section provides an overview of the three research papers with regard to purpose, research design, authorship, submission details, current status, and conference presentations.

Paper 1 | Intraday Online Information Demand and its Relationship with REIT Prices

This study analyzes the intraday information demand of internet users and its relationship with US REIT prices. For this purpose, trading strategies based on hourly changes in search volumes, provided by Google Trends, are identified and compared to buy-and-hold strategies of the underlying REITs. Moreover, it is investigated in which market phase and before which trading signal, Google trading strategies are more successful in predicting intraday REIT price movements. The results are validated by including the stocks of the DJIA index as a control group.

Authors: Katrin Kandlbinder, Marian Alexander Dietzel

Submission to: Journal of Real Estate Portfolio Management

Current Status: Under Review

This paper was presented at the PhD-Session of the 2016 Annual Conference of the American Real Estate Society (ARES) in Denver, US and at the 2016 Annual Conference of the European Real Estate Society (ERES) in Regensburg, Germany.

Paper 2 | Leveling the Playing Field:
Out-of-Town Buyer Premiums in US Housing Markets Over Time

The main purpose of this paper is to investigate whether out-of-town buyers do in fact pay higher prices for real estate and why, and whether this premium decreased from 2005 to 2015 due to better information availability. Using a sample of 15,795 condominium transactions in Miami Dade County, a hedonic regression model is developed and extended by out-of-town, anchoring and wealth variables. By applying a sophisticated statistical matching technique, namely propensity score matching, the robustness of the results is ensured, and selection bias avoided.

Authors: Katrin Kandlbinder, Norman G. Miller, Michael Sklarz

Submission to: International Journal of Housing Markets and Analysis

Current Status: Forthcoming

This paper was presented at the 2017 Annual Conference of the American Real Estate Society (ARES) in San Diego, US and at the PhD Session of the 2017 Annual Conference of the European Real Estate Society (ERES) in Delft, the Netherlands.

Paper 3 | Predicting Real Estate Market Movements:
the First Textual Analysis-Based-Sentiment Application in Germany

By applying a dictionary-based approach to German real estate newspaper articles, the purpose of this paper is to determine whether there is a relationship between different sentiment measures and German housing prices. Generating the first *German Real Estate Sentiment Dictionary* with 14,137 objectively validated words, enables extracting sentiment from 125,462 newspaper articles published by the *Immobilien Zeitung* – the major German real estate news provider. A vector autoregressive framework and out-of-sample forecasts are then utilized to examine the dynamic relationship between news-based sentiment measures and the German housing market from 2007 to 2017.

Authors: Jessica Ruscheinsky, Katrin Kandlbinder, Wolfgang Schäfers, Marian Alexander Dietzel, Karim Rochdi

Submission to: Journal of European Real Estate Research

Current Status: Under Review

This paper was presented at the 2018 Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, US.

1.4 References

Antweiler, W. and Frank, M.Z. (2004), “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards”, *The Journal of Finance*, Vol. 59 No. 3, pp. 1259–1294.

Askitas, N. and Zimmermann, K.F. (2009), “Google Econometrics and Unemployment Forecasting”, *Applied Economics Quarterly*, Vol. 55 No. 2, pp. 107–120.

Baker, M. and Wurgler, J. (2007), “Investor Sentiment in the Stock Market”, *Journal of Economic Perspectives*, Vol. 21 No. 2, pp. 129–151.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), “Commercial Real Estate Valuation. Fundamentals Versus Investor Sentiment”, *The Journal of Real Estate Finance and Economics*, Vol. 38 No. 1, pp. 5–37.

Da, Z., Engelberg, J. and Gao, P. (2014), “The Sum of All FEARS Investor Sentiment and Asset Prices”, *Review of Financial Studies*, Vol. 28 No. 1, pp. 1–32.

Dietzel, M.A., Braun, N. and Schäfers, W. (2014), “Sentiment-based commercial real estate forecasting with Google search volume data”, *Journal of Property Investment & Finance*, Vol. 32 No. 6, pp. 540–569.

Fama, E.F. (1970), “Efficient Capital Markets: A Review of Theory and Empirical Work”, *The Journal of Finance*, Vol. 25 No. 2, pp. 383–417.

Gentzkow, M. and Shapiro, J.M. (2010), “What Drives Media Slant? Evidence From U.S. Daily Newspapers”, *Econometrica*, Vol. 78 No. 1, pp. 35–71.

Hohenstatt, R., Käsbauer, M. and Schäfers, W. (2011), “"Geco" and its Potential for Real Estate Research: Evidence from the U.S. Housing Market”, *Journal of Real Estate Research*, Vol. 33 No. 4, pp. 471–506.

Lüdering, J. and Tillmann, P. (2016), “Monetary Policy on Twitter and its Effect on Asset Prices: Evidence from Computational Text Analysis”, *MAGKS Paper on Economics*, No. 12, pp. 1–34.

Mori, M. (2015), “Information Diffusion in the U.S. Real Estate Investment Trust Market”, *The Journal of Real Estate Finance and Economics*, Vol. 51 No. 2, pp. 190–214.

Vlastakis, N. and Markellos, R.N. (2012), “Information demand and stock market volatility”, *Journal of Banking & Finance*, Vol. 36 No. 6, pp. 1808–1821.

2 Intraday Online Information Demand and its Relationship with REIT Prices

Abstract

A fictional trading strategy based on hourly Google search volumes is developed for the MSCI US REIT Index to show whether there is a relationship between intraday online search interest and REIT market movements. Furthermore, we investigate in which market circumstances this trading strategy has the best predicting abilities and we examine the controversial questions of correlation and causality between search volumes and prices. The results indicate that search volumes indeed have the ability to predict intraday REIT market movements, as the Google trading strategy achieves an outperformance of 7.37 percentage points on average, compared to a buy-and-hold strategy of the underlying REIT. In falling market phases the performance results of the Google trading strategy are substantially better than in rising market phases. On average, there is a statistically significant correlation of -0.11 and a causality flow from prices to search volumes. The findings yield new insights into the information-gathering behavior and are therefore useful for understanding and anticipating the relationship between market participants' information demand and REIT price movements.

2.1 Introduction

In the 1970ies, Eugene Fama stated in his efficient market hypothesis that market prices and stock returns reflect all available information (Fama, 1970). By now, however, most researchers argue that prices are also influenced substantially by investor sentiment. Up to this point, many studies have made efforts to explain why sentiment exists, where it comes from, and how it is created (e.g. Rosenberg *et al.*, 1985; DeLong *et al.*, 1990; Baker and Wurgler, 2007). In line with Black (1986) and Barberis *et al.* (1998), we hypothesize that sentiment is created from incoming information. Either a market participant is actively seeking information or is being informed. He then uses his individual information set to form an opinion about market developments, stock prices, etc. and will behave accordingly. Therefore, information is supposed to be the most valuable and highly sought-after asset in financial markets (Vlastakis and Markellos, 2012).

The most important source for acquiring information nowadays is the internet. In our digitized society, a steadily increasing number of internet users visit websites of search engines every day, as they act as a gateway to information. Each query request can be seen as an individual "vote", because we leave information about our interests codified as search terms. By incorporating Google search volumes as a proxy for investor sentiment, several researchers have already shown that there is a relationship between search volumes and general stock returns. They conclude that search volumes are a direct and unambiguous measure of attention and therefore have the potential to reveal sentiment (Bank *et al.*, 2011; Da *et al.*, 2011; Preis *et al.*, 2013; Curme *et al.*, 2014; Da *et al.*, 2014).

Although the relationship between Google search volumes and stock price movements in the general stock market has been documented in literature, the possible relationship and predicting abilities of Google search volumes in context of Real Estate Investment Trusts (REITs) has not been addressed by many researchers. Even though it has been shown that sentiment evidently plays an important role in REIT pricing as well (Lin *et al.*, 2009). In this research, we focus on information procurement, namely using search volume data provided by Google Trends[1] as a proxy for investor sentiment, in order to investigate the following research questions.

Firstly, do search volumes provided by Google Trends have the ability to successfully predict intraday REIT price movements? Secondly, in which market phases (falling, stagnant, rising) and before which trading signal (buy or sell) does information procurement have the best prediction abilities for REITs? Thirdly, we concentrate on the much debated question of correlation and causality between search volumes and prices.

[1] Google Trends is a public tool, which offers search volume indices for all kinds of search queries.

The application to the REIT market suggests itself as it has already been shown that REIT prices are affected by sentiment just as general stock prices (Clayton and MacKinnon, 2001; Lin *et al.*, 2009; Das *et al.*, 2015b). Furthermore, REITs as a special asset class have unique characteristics that can lead to various advantages when establishing a trading strategy based on search volumes compared to general stocks. REITs are characterized by high homogeneity amongst their assets. Every REIT is obliged to have a high exposure to real estate, whereas a bundle of general stocks like the Dow Jones Industrial Average (DJIA) is assigned to nine different industries. Furthermore, REIT investors are supposed to be more qualified and less diverse compared to the average DJIA investor. REIT investors already know about the specifics of this particular investment vehicle and formulate their search queries accordingly. Furthermore, the market capitalization is smaller and the average number of investors is lower compared to blue chip stocks. Therefore, REITs supposedly capture a lower diversity of opinions and less noise within the search queries. Hence, with these characteristics, we assume that a trading strategy based on search queries for REITs will achieve better trading results, compared to DJIA stocks. In order to test this hypothesis we apply the same methodology for both REITs and DJIA stocks.

We base our analysis on a methodology introduced by Preis *et al.* (2013), who incorporate trading strategies based on changes in Google search volumes to show that these trading strategies achieve greater profits than a random buy-and-hold strategy. As a dataset we use hourly search volumes and stock prices of the Top 20 MSCI US REITs over 5 months, from November 2015 to March 2016.

This paper makes a research contribution by expanding previous research in the following ways. First, this is the first paper which uses intraday search volumes and prices in order to gain more accurate and detailed insights into information procurement behavior. This knowledge is very useful to improve trading strategies and price prediction, thus reducing the measurement imprecision of weekly data and capturing short-term sentiment fluctuations. All Google Trends research so far focuses on weekly or monthly trading frequencies, but as Hu *et al.* (2015) state, investor sentiment may vary within a short time frame. Canbaş and Kandır (2009) indicate that intraday data allow for a more reliable and efficient estimation of the effect of sentiment factors on stock prices. Second, we concentrate on REIT price movements and the relationship with search volumes, a previously almost neglected asset class in Google sentiment literature. Finally, we test the validity of our results by including the 30 DJIA stocks as a control group in order to investigate which asset class performs better with Google trading strategies.

Our results show that trading strategies for REITs based on intraday search volumes have the potential to outperform a simple buy-and hold strategy by 7.37 percentage points on average. In falling market phases the performance results of the Google trading strategy are substantially better, than in rising market phases. On average, we find a statistically significant negative correlation of -0.11 and a causality flow from prices to search volume.

The remainder of the paper is structured as follows. The next section reviews the relevant literature. Section 3 discusses the datasets for measuring investor attention/sentiment and determining where the capital market data is obtained. Section 4 outlines the theoretical and methodological background, with Section 5 presenting the results of the empirical application. The final section concludes the paper.

2.2 Literature Review

2.2.1 Google Search Volume

The internet has become a central source of information for day-to-day decisions. As most information-gathering now takes place online, search data has the unique potential to objectively and directly reveal the underlying beliefs of an entire population. Therefore, it is a powerful measure of attention.

A growing number of researchers has employed Google search volume data in different research fields and has shown that search engine queries for specific keywords can be linked to a variety of real-world events. Ginsberg *et al.* (2009) were among of the first to use Google search volumes in the field of epidemiology to identify and predict influenza "hot spots" in the US. In the field of economics, later on the same year, there were the first applications by Choi and Varian (2009), who predicted values of economic indicators on this basis. Askitas and Zimmermann (2009) created an index based on search words that job seekers use to find a job, in order to forecast unemployment rates. Further economics-related research on Google search volumes was conducted by Guzman (2011), McLaren and Shanbhogue (2011), and Dzielinski (2012).

Kahneman and Tversky (1979) have already suggested in their Prospect Theory that investor psychology and sentiment play a crucial role in return generation. Therefore, it is not surprising that Google search volumes, which serve as a proxy for investor attention or sentiment, have been applied in the field of finance as well. Da *et al.* (2011) make use of Google Trends to construct a new measure of investor attention to predict trading volume and stock market returns for the Russel 3000 stocks. Furthermore, by using a vector autoregression (VAR) framework, they conclude that internet-based search volume indices capture investor attention more efficiently than commonly used attention measures.[2] Similarly, Drake *et al.* (2011) employ search queries to quantify investor demand by using company-related information as tickers of S&P 500 stocks. Following a different approach, Da *et al.* (2014) construct a Financial and Economic Attitudes

[2] Indirect proxies for investor attention are for example extreme returns, trading volume, news and headlines, advertising expenses, and price limits.

revealed by Search (FEARS) index, by aggregating daily search volume-indices for non-company keywords related to household, financial, and economic concerns. Vlastakis and Markellos (2012) employ company names as search terms to approximate information demand and public interest at the firm- and market-level for the 30 largest stocks traded on NYSE. The results indicate that information demand is positively correlated with volatility and with trading volume.

In contrast to the positive correlation, Preis *et al.* (2010), Joseph *et al.* (2011), Preis *et al.* (2013), Curme *et al.* (2014), and Da *et al.* (2014) find evidence of a negative relationship between internet searches and subsequent stock market movements. Preis *et al.* (2013), for example, demonstrate that enormous increases in DJIA stock prices were preceded by a decrease in financially related search volumes like “debt” and vice versa. Furthermore, they implement a search query-based trading strategy which generated significantly higher returns than the benchmark.

Most empirical research has focused on the capital market. However, Google Trends research has also been conducted in the field of direct real estate. Wu and Brynjolfsson (2009), Hohenstatt *et al.* (2011), Dietzel *et al.* (2014), Hohenstatt and Käsbauer (2014), and Das *et al.* (2015a) confirm the forecasting abilities of Google Trends for the property market, both for housing and commercial real estate.

2.2.2 Investor Sentiment and REIT Pricing

The most widely known theory on the role of investor sentiment is that of DeLong *et al.* (1990), which demonstrates that investors are subject to sentiment and that they trade on non-fundamental information as well. Barkham and Ward (1999) were one of the first to investigate the role of investor sentiment within the securitized real estate sector. They conclude that market-wide sentiment is just as influential as specific company factors in explaining the discount or premium to NAV in UK property companies. Clayton and MacKinnon (2003) find that investor sentiment is important to REIT pricing even after accounting for REIT and private market liquidity. Lin *et al.* (2009) confirm the significant influence of investor sentiment on REIT returns. They state that when investors are optimistic, REIT returns become higher and vice versa. The results are even robust when conventional control variables are considered. Chiang and Lee (2010) answer the question which kind of REITs are more prone to sentiment by using correlated trading as a proxy for sentiment. They find that sentiment is stronger for illiquid REITs that appear to be preferred by individual investors. In a more recent paper, Ro and Gallimore (2014) investigate herding behavior as a form of sentiment for REITs and real estate mutual funds. They support the view that REITs are relatively more transparent but herding behavior still exists.

So far, there are two studies relating to the application of search volume data provided by Google Trends and the relationship with REIT pricing. Rochdi and Dietzel (2015) show that there is a positive relationship between asset-specific online search interest and price movements in the US

REIT market. By establishing an investment strategy based on weekly changes in Google search volume, they find that real estate-related search terms are more suitable than general terms for predicting REIT market movements. Das *et al.* (2015b) use quarterly data for 21 US MSAs to identify a connection between increased online searches and higher REIT returns. They find some evidence that the searches are fundamentally associated with REIT returns in the short run.

Apart from this literature focusing on weekly or quarterly search volumes, there are no research studies testing the relationship between REITs and search volumes in detail by using intraday data. Inspired by Preis *et al.* (2013) and Rochdi and Dietzel (2015), this present study aims to fill this research gap by dissecting the information demand-price relationship in the US REIT market. In particular, we test whether intraday search-volume based trading strategies for individual REITs outperform a traditional buy-and-hold strategy and, most importantly, under which market circumstances the outperformance is generated.

2.3 Data

2.3.1 Google Search Volume Data

Google, the search engine with the highest market share in the US, offers a publicly available search volume index for all kinds of search of queries. Data are made public via the tool Google Trends.[3] Until 2015, it was only possible to download search volumes on a weekly basis, starting in January 2004. Since June 2015, Google Trends also makes available search query data on an intraday basis. This means that search-interest logs can be traced close to real-time, every hour or even every minute.

However, due to the very large data packages, Google Trends provides hourly data only for the previous week. The finer the data granularity, the shorter the search volume history. Due to very laborious week-by-week downloads, we limit the sample of hourly search volume data to a period ranging from November 2015 to March 2016, as this timeframe is highly representative with falling, rising, and stagnant markets.[4] To investigate the relationship between hourly search volume data and stock prices, the total sample of search query data, which are available 24 hours a day, has to be adjusted to the trading hours of the New York Stock Exchange. Therefore, we generate a sample, with 78 trading days with seven regular trading hours a day, and two days with

[3] Available at: *https://www.google.com/trends/*.

[4] Due to the fairly short time frame, we ignore certain difficulties that arise when working with weekly data. For example, when doing research with terms whose relevance has increased tremendously over time, weekly data will be valued as zero, especially in the starting years of Google Trends data (2004, 2005), due to the substantial change in volume and the normalization procedure. In our study, the popularity of the search terms is considered to be constant over slightly more than five months and must not be adjusted.

four trading hours a day after holiday, within the period of observation. This results in a total sample of 554 hourly observations.

In addition to the timely data-frequency, Google Trends offers some filtering functionalities such as location and search category. Employing various restrictive (sub-) categories can imply that there may not be enough search traffic for a specific search term. For this reason, we decide to follow Bank *et al.* (2011), Preis *et al.* (2013), and Curme *et al.* (2014), who do not apply any category at all to capture the maximum number of relevant search volumes.

As both indices in this paper (MSCI US REIT Index and DJIA) are related to the US and traded on the New York Stock Exchange, we limit the search volume results to the US, following Curme *et al.* (2014) and Da *et al.* (2014). Preis *et al.* (2013) note that it is widely recognized that investors tend to trade mainly in their own domestic market. Therefore, search data from US users only are intended to capture information-gathering behavior more precisely than that of worldwide Google users.

2.3.2 Google Search Terms

According to Bank *et al.* (2011), Da *et al.* (2011), and Vlastakis and Markellos (2012), there are two main methods for employing company-related search terms, namely company names or stock tickers.

We use the company name plus the word "REIT" (e.g. Boston Properties+REIT) as a search term. Since we are interested in the impact of investor attention on trading and asset pricing, we aim to capture only the group of people who are interested in financial information about a specific share. Due to low search volumes for REITs, the quantity of observed REITs had to be narrowed down to the Top 20 of the MSCI US REIT Index measured by market capitalization. Smaller REITs with lower market capitalization are likely to be more unknown und therefore deliver useless search volume as it is nearly zero. Hence, a combination with the word "REIT" ensures that there will be enough search traffic and avoids capturing fuzzy searches for the Top 20.

For our control group, the 30 DJIA stocks, we focus on the company name as well[5], combined with the word "stock" (e.g. Apple stock). Combinations with other finance-related words like "share" deliver not as much search volume as stock.

[5] We decide to use names instead of tickers for two reasons. First, when using tickers, one has to be careful of ambiguous meanings. CAT, for example, is the ticker symbol for Caterpillar, but can easily be confused with the animal. Second, with the firms' names, we expect the search volume index to capture a much broader and more relevant audience, because it seems unlikely that non-institutional internet users would search for a company by using its stock ticker symbol.

2.3.3 Data Issues

Google Trends provides search volumes, using a finite integer scale from zero (which yields the lowest search volume) to 100 (which represents the highest value), instead of reporting the raw quantities of searchers. That means that the chart for the same search term changes, as soon as a new maximum or minimum has been reached. Bank *et al.* (2011) argue that this normalization has its pros and cons. On the one hand, this transformation done by Google, eliminates the trend towards a growing number of search queries, due to higher internet use, but on the other hand, it prevents us from taking advantage of the absolute numbers of search volumes.

Furthermore, search volume data change slightly over time, due to Google's extraction procedure and data normalization. Choi and Varian (2009) address this problem of sampling noise in one of the first studies about Google Trends data. This inconsistency becomes obvious when data is downloaded for the same time range, but on different occasions. In our case, the most important fact for the trading strategy is the accordance of the trading signal on whether to buy or sell (0 or 1) and not the absolute numbers of search volume from different downloads. In order to find out whether there is a crucial discrepancy, we follow Da *et al.* (2011), Preis *et al.* (2013), and Da *et al.* (2014) and test the correlation between the search volumes for three independent data requests every two days within one week. The result is an over 98% consistency in buy or sell signals. Therefore, we believe that the impact of this sampling error is small and should not bias the results.

Descriptive statistics of the search volume for the Top 20 MSCI US REITs can be found in Appendix 2.1.

2.3.4 Capital Market Data

The capital market data are derived from the Yahoo Finance chart API. The prices have to be downloaded every two weeks, as the data history covers only the last 10 trading days. They include the timestamp, opening and closing price, and the trading volume of the single stocks on an hourly basis. Hence, weekends, holidays, etc. had to be accounted for. Thus, the total number of observations results from the trading hours of the New York Stock Exchange. The New York Stock Exchange is open from Monday through Friday 9:30 a.m. to 4:00 p.m. Since we have an hourly setting, our sample contains 7 trading hours from 10:00 a.m. to 4:00 p.m. on five trading days per week, which results in 35 observations per week over 17 weeks. Furthermore, we had to account for holidays like Thanksgiving Day, Christmas, New Year's Day, etc. This reduces the number of observations from 595 to 554, as the NYSE is closed during these times, and consequently no REIT prices are available.

2.4 Methodology

In order to find out whether search volumes provided by Google Trends have the ability to successfully predict intraday REIT price movements, we apply a methodology similar to that of Preis *et al.* (2013) and implement a hypothetical trading strategy. The trading strategy is based on the relative changes in search volumes (SV) and quantifies the changes in information-gathering as follows:

$$\Delta SV_{(hour_{t,T})} = SV_{(hour_t)} - SV_{(hour_{t-1,T})}$$
$$with: \; SV_{(hour_{t-1,T})} = \frac{1}{T}\sum_{i=1}^{T} SV_{(hour_{t-i})} \qquad (1)$$

Notes: Where t is measured in units of hours and T = 5. The relative changes in search volumes is the basis for the hypothetical trading strategy. The trading signal (buy or sell) or trading rule itself is determined in a second step.

Comparing only the change in search volume of the actual search volume in *t* with the search volume in *t-1* would induce too much noise and therefore bias the trading signals. In order to ensure the robustness of our results, the relative changes in search volume ($\Delta SV_{(hour\ t,T)}$) of a specific search term are determined as the mean value over the search volume of the five preceding hours for *T= 5* hours, following Preis *et al.* (2013). Furthermore, we assume that there is a time gap between the research process and the final transaction for considering or collecting further information about an explicit stock. This implies that high search volume in *t-4*, for example, can also affect price movements in *t*.

By means of the trading strategy, we aim to anticipate intraday REIT price movements. Of course, profits can only be made if the trading strategy predicts the REIT price movement correctly, in particular around significant movements. The trading signal is derived from the relative changes in search volumes ($\Delta SV_{(hour\ t,T)}$) and is formulated as follows:

$$Trading\ Signal_{(hour_t,pos)} \begin{cases} 0 = buy, & if\ \Delta SV_{(hour_{t,T})} > 0 \\ 1 = sell, & if\ \Delta SV_{(hour_{t,T})} < 0 \end{cases} \qquad (2)$$

Notes: Here, 0 is defined as a buy signal and 1 as a sell signal. A positive value of $\Delta SV_{(hour\ t,T)}$ indicates an upward trend. Therefore, we take a long position and buy the REIT. However, if $\Delta SV_{(hour\ t,T)}$ is lower than zero, we expect the market to fall, take a short position and sell the REIT.

Up to this point, it is unclear whether an increase in search volume is directly related to a subsequent increase or decrease in the stock price. There is a rather controversial body of literature regarding the direction of correlation (e.g. Da *et al.*, 2011; Joseph *et al.*, 2011; Preis *et al.*, 2013). Barber and Odean (2008) and Da *et al.* (2011) argue that individual investors are net buyers of "attention-grabbing stocks" and assume a positive correlation. They suggest that searching for a stock online is relatively more useful for somebody considering buying a stock rather than selling it. Someone who is willing to buy wants to collect information about the company history and recent stock performance, so as to narrow down the number of viable alternatives. Whereas a person who owns the stock is already presumably knowledgeable about the stock. By contrast, Joseph *et al.* (2011), Preis *et al.* (2013), and Curme *et al.* (2014) indicate that people tend to gather more information online in times of uncertainty and concern. They conclude that a high level of interest in certain stocks predicts temporary downward market pressure and therefore pursue a negative correlation.

We aim to derive a more objective and detailed approach to determine the direction of the relationship between search volumes and REITs. We use company names as search terms which are absolutely objective. The search term "debt", for example, as used by Preis *et al.* (2013), induces a negative bias upfront. Furthermore, we employ an intraday setting and concentrate on individual REITs, not on indices, when calculating the correlation between search volumes and prices.

Another aim of this study is to investigate possible causal relationships between information demand and price movements. We undertake a pairwise Granger Causality analysis in order to find out whether prices are driven by search volumes or search volumes are affected by price movements. As a specification, we use the Akaike Information Criterion (AIC) to identify the optimal lag order.

Basic model for testing Granger Causality:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \beta_1 z_{t-1} + \alpha_2 y_{t-2} + \beta_2 z_{t-2} + \cdots \tag{3}$$

$$z_t = \omega_0 + \vartheta_1 y_{t-1} + \sigma_1 z_{t-1} + \vartheta_2 y_{t-2} + \sigma_2 z_{t-2} + \cdots \tag{4}$$

Notes: For Granger Causality analysis the variables used in the model have to be stationary. Therefore, we take the first difference of the prices. y is defined as the first difference of the REIT prices and z is the search volume over time.

Equation (3) allows us to test whether past values of search volumes (z) help to forecast prices (y), after controlling for past values of y, whereas equation (4) indicates whether past values of REIT prices (y) help to forecast search volumes (z).

2.5 Results

2.5.1 REIT Trading Strategy Results

In order to anticipate stock market movements, we wish to investigate whether Google Trends yields useful real-time insights into the information demand of traders. To quantify the prediction abilities and the quality of the fictional trading strategy, we introduce three different measurements.

Firstly, ‘outperformance’ which is calculated by the performance of the trading strategy minus the performance of the simple buy-and-hold strategy of the underlying REIT. The result in Exhibit 2.1 shows that the Google trading strategy of 20 MSCI US REITs achieves an outperformance of 7.37 percentage points on average from November 2015 to March 2016. If viewed individually, *Equity Residential* generates the highest outperformance over the time of consideration with 45.60 percentage points. 15 out of 20 REITS do report a positive outperformance. This means that the trading strategy based on search volume changes would have created higher returns than an ordinary buy-and-hold strategy.

As a second measure for quantifying the prediction abilities of the Google trading strategy, we calculate ‘hit rates’, which are defined as the number of a particular strategy’s correct predictions divided by its total number of predictions. Bearing in mind that profits can only be made if the trading strategy anticipates the right stock price movement, the hit rate is an important quality measure of the trading strategy. The hit rate for the overall sample is fairly moderate at 50.69% but statistically significant at the 10% level.

One would assume that high hit rates result in high returns of the trading strategy. However, in line with Joseph *et al.* (2011), the results show that the REIT with the highest hit rate is not automatically the best performer. As presented in Exhibit 2.1, *Avalonbay Communities*, for example, has the highest hit rate with 54.61% but reports only an outperformance of 6.23 percentage points.

Since the period of consideration is characterized by large market movements, we extend the hit rate by the third trading strategy quality measure which we refer to as an ‘abnormal hit rate’. Hence, we measure the correct prediction of the 10% highest positive and negative price movements of the underlying REIT. The abnormal hit rate for our overall sample of 20 REITs is 68.48% on average and statistically significant at a 1% level.

Exhibit 2.1 | Trading strategy results

MSCI US REITs	Outperformance	Hit Rate	Abnormal Hit Rate	Correlation	t-statistic
Equity Residential	45.60	52.26%	80.25%	-0.18***	-4.20
Equinix	36.13	50.45%	75.00%	-0.16***	-3.84
Host Hotels & Resorts	20.83	50.09%	61.36%	-0.15***	-3.64
HCP	18.25	50.81%	69.31%	0.02	0.57
Boston Properties	13.27	50.27%	69.01%	-0.23***	-5.58
SL Green Realty Corp	12.55	47.92%	65.43%	-0.19***	-4.68
Federal Realty Inv Trust	11.40	51.72%	68.85%	0.31***	-7.58
Ventas	10.63	51.54%	70.65%	-0.35***	-0.09
ProLogis	10.11	51.54%	71.95%	-0.20***	-4.86
Macerich	7.63	48.82%	73.33%	-0.01	-0.22
Avalonbay Communities	6.23	54.61%	67.53%	-0.07	-1.55
Simon Property Group	5.46	49.19%	66.20%	-0.04	-0.94
Vornado Realty Trust	2.76	51.36%	60.29%	-0.16***	-3.73
Public Storage	1.17	52.26%	80.26%	0.24***	5.85
Digital Realty Trust	0.61	53.16%	67.61%	0.14***	3.40
Essex Property Trust	-0.11	51.18%	69.74%	-0.14***	-3.34
General Growth Properties	-6.11	46.47%	67.37%	0.08	1.88
Kimco Realty Corp	-9.50	47.56%	63.89%	0.34***	8.40
Welltower	-15.64	51.18%	60.00%	0.09	2.04
Realty Income Corp	-23.81	51.36%	61.54%	0.25***	6.12
MSCI US REITs	**7.37**	**50.69%**	**68.48%**	**-0.11*****	**-2.55**

Notes: This exhibit depicts the main performance measures, outperformance, hit rate, and abnormal hit rate and the Pearson correlation coefficient with t-statistics, ranked by outperformance. The outperformance is defined as return of the Google trading strategy minus the return of the benchmark of the underlying stock measured in percentage points. The abnormal hit rate is the hit rate for the 10% highest positive and negative price movements and therefore represents the most volatile market phases of the 20 MSCI US REITs. The correlation indicates the direction of the relationship between Google search volume and prices. * Indicates significance at the 10% level. ** Indicates significance at the 5% level. *** Indicates significance at the 1% level.

We find that the best performers tend to have higher abnormal hit rates (70 to 80%) during volatile market phases. This suggests that the correct prediction of big jumps with abnormal returns (abnormal hit rate) is much more important than the absolute prediction accuracy (hit rate). These results point in the same direction as the findings of Vlastakis and Markellos (2012), Curme *et al.* (2014), and Rochdi and Dietzel (2015), who state that people tend to have a higher information demand in times of higher volatility and increased uncertainty about future market developments.

Generally, the findings suggest that information-demand-based trading strategies have the potential to outperform the benchmark in most cases. The overall performance results – outperformance, hit rate and abnormal hit rate – provide evidence that Google is indeed used by short-term traders as an information source when making investment decisions, given that Google trading strategies based on search volume outperform the benchmark. Therefore, the movement of Google search volumes obviously includes valuable information in order to anticipate REIT market movements. This important finding for the real estate sector is in line with Preis *et al.* (2013) and Curme *et al.* (2014), who state that changes in the search activity of Google users give an indication of financial market movements.

We also address the controversial issue of correlation between Google search volumes and REIT prices. On average, we find a slightly negative correlation of -0.11, which is highly statistically significant at the 1% level. Therefore, we support the hypotheses of Joseph *et al.* (2011) and Curme *et al.* (2014) that high Google search volumes are accompanied by price declines. Individually, the picture of correlation is not as clear as on an aggregated level. Overall, 60% of the REITs report a negative correlation whereas 25% show a statistically significant positive correlation. Hence, we discover an interesting finding: For REITs with a negative correlation the outperformance is substantially high and always positive. If REIT prices and volumes are positively correlated, the outperformance tends to be negative.

2.5.2 Performance Measures in Detail

In order to gain a better understanding of how investors use Google as an information source before an investment, we separate the Google trading strategy into the two trading signals (buy/sell) and divide the time frame into certain sub-periods of falling, rising, and stagnant market phases.

Grullon *et al.* (2004), Barber and Odean (2008), and Da *et al.* (2011) argue that investors tend to use the internet more often for gathering company-related information before purchasing a stock, rather than selling it. In order to test whether this phenomenon is persistent, the trading signals are divided into long-only and short-only signals. A higher hit rate for long (short) trading signals indicates that information demand before buying (selling) is more representative, as the Google trading strategy will report better performance measures.

Furthermore, the reason for dividing the time frame into market phases is that García (2013) finds evidence that investors tend to use different decision-making rules in recessions than in expansions. Consequently, information-gathering behavior, which is depicted by search volumes, is supposed to vary according to different market phases. As the Google trading strategy is based on relative changes of search volumes, it is necessary to take a closer look at the performance results in the context of varying market phases.

Exhibit 2.2 shows, that the outperformance is 14.06 percentage points higher in falling markets, than in rising markets. We believe this is because investors investigate information more intensely in a situation of increased uncertainty and thereby reveal clearer signs of their investment behavior. Furthermore, the results support the findings of Joseph *et al.* (2011) and Vlastakis and Markellos (2012), who point out that the effect of Google information demand increases during downward market phases.

Exhibit 2.2 | Performance results for different market phases divided into long and short signals

Market phase	Outperformance	Hit rate			Abnormal Hit Rate		
		long		short	long		short
falling	8.39	40.39%	<	68.15%	**79.03%**	>	35.29%
stagnant	2.84	52.17%		57.40%	**69.12%**	>	38.92%
rising	-5.67	52.70%	>	43.53%	**62.36%**	>	38.34%

Notes: The exhibit provides Google trading strategy performance measures – outperformance, hit rate, and abnormal hit rate – for falling, stagnant, and rising markets, divided into long and short signals.

Concerning the hit rates, we detect interesting patterns. In falling markets, the hit rate for short signals exceeds the rate for long signals, whereas in rising markets, we observe an inverse relationship. This means that in falling market phases, Google trading strategy short signals have greater potential to anticipate the correct REIT price movement and in rising markets, long signals. This phenomenon can be explained by the fact that in falling (rising) markets, the number of short (long) signals is essentially higher, and therefore, the hit rate should be higher as well.

The pattern changes when looking at the abnormal hit rate results which paint a clear picture. Irrespective of the market phase, the abnormal hit rate for long signals is always substantially higher (62-79%) than for short signals (35-38%). In highly volatile market phases, information demand seems to be an unambiguous indicator of price changes. Especially in falling markets, both the outperformance and the abnormal hit rate performed very strongly. On the one hand, this is due to large market movements during volatile phases and therefore higher possible returns. On the other hand, it suggests that in times of uncertainty, when internet users have an increased appetite for information, Google trading strategies work exceptionally well in anticipating upwards movements (buy signals) in falling markets or so called turning points. Additionally, we find an unambiguous positive correlation of nearly 60% between abnormal hit rate long and outperformance. This finding somewhat suggests itself. The higher the abnormal hit rate “long”, the better is the prediction ability for abnormal positive returns and therefore finally, the higher the outperformance.

2.5.3 Granger Causality

A clear causal relationship between information demand and prices can provide insights as to whether search volumes drive prices or vice versa. If, for example, information demand Granger-causes REIT prices, this would suggest that search volumes (SV), as a proxy for sentiment, cause price movements. Whereas, if the opposite is true, this may mean that people react to the price changes by further investigating on the internet.

The results, shown in Exhibit 2.3, suggest that, on average, there exists significant causality flowing from prices to search volumes with a p-Value of 0.0679. Individually, we find a statistically significant causal relationship from prices to search volumes for six REITs. The calculated statistics for the reverse causality from search volumes to prices are much smaller, but achieve statistical significance in four cases. Therefore, almost half of the REIT sample show robust evidence of significant causality between search volume and prices in one way or another. Darrat *et al.* (2003) and Vlastakis and Markellos (2012) come to the same conclusions regarding the Granger Causality application to the general stock market. The results of bidirectional causality do not offer themselves a clear interpretation. A plausible explanation for the bidirectional causality is that the two variables interact with each other. For example, when the information, which people are seeking for online, is already incorporated in the prices. Then prices drive search volumes to some extent. But if prices do not reflect these information, search volume drives prices, effectively causing the two variables to dynamically interact.

Exhibit 2.3 | Granger Causality analysis

		SV Granger cause prices		**Prices Granger cause SV**	
MSCI US REITs	**AIC (lags)**	**p-Value**	**F Value**	**p-Value**	**F Value**
Avalonbay Communities	9	0.5715	0.8482	0.3046	1.1771
Boston Properties	9	0.2381	1.2865	0.848	0.5378
Digital Realty Trust	15	0.1821	1.3248	0.3801	1.0718
Equinix	3	**0.0916***	2.1277	0.9163	0.1706
Equity Residential	9	0.8959	0.4696	0.3589	1.0997
Essex Property Trust	9	0.6613	0.752	0.1637	1.4418
Federal Realty Inv Trust	15	0.2058	1.2862	**0.0017*****	2.4577
General Growth Properties	15	0.7548	0.7313	**0.0591***	1.6244
Welltower	9	0.5191	0.9055	0.5443	0.8777
HCP	16	**0.0970***	1.5792	0.3581	1.0896
Host Hotels & Resorts	15	0.2147	1.2653	**0.0343****	1.7581
Kimco Realty Corp	16	**0.0808***	1.5259	**0.0711***	1.5578
Macerich	9	0.1407	1.5014	0.6124	0.8042
Realty Income Corp	19	0.8947	0.6202	**0.0422*****	1.6223
ProLogis	15	0.6725	0.8058	**0.0320*****	1.7746
Public Storage	3	**0.0870***	2.1897	0.1673	1.6876
SL Green Realty Corp	15	0.2783	1.1806	0.1797	1.3201
Simon Property Group	8	0.3832	1.0666	0.4463	0.9837
Vornado Realty Trust	9	0.3177	1.1575	0.469	0.9625
Ventas	3	0.9195	0.1657	0.2006	1.5447
MSCI US REITs	**15**	**0.2268**	**1.2479**	**0.0679***	**1.5893**

Notes: The exhibit depicts the results of the pairwise Granger Causality test. We use the first difference of the prices to generate stationary variables for the Granger Causality analysis. The optimal lag length was identified by the Akaike Information Criterion (AIC). * Indicates significance at the 10% level. ** Indicates significance at the 5% level. *** Indicates significance at the 1% level.

2.5.4 REITs vs. DJIA Stocks

In order to investigate which asset class (REITs or stocks) perform better with the Google trading strategy and to test the validity of our methodology, we include the 30 DJIA stocks as a control group.

Panel A in Exhibit 2.4 shows the performance measures outperformance, hit rate, and abnormal hit rate on average for both REITs and DJIA stocks. For all three performance measures, the REITs achieve better trading strategy results. The outperformance, hit rate, and abnormal hit rate for REITs are higher, compared to DJIA stocks.

Exhibit 2.4 | REITs vs. DJIA stocks

Panel A

Index	**Outperformance**	**Hit Rate**	**Abnormal Hit Rate**
REITs	7.37	50.69%	68.48%
DJIA	6.03	50.23%	55.74%

Panel B - REITs

Market phase	**Outperformance**	**Hit Rate**			**Abnormal Hit Rate**		
		long		**short**	**long**		**short**
falling	8.39	40.39%	<	68.15%	**79.03%**	>	35.29%
stagnant	2.84	52.17%		57.40%	**69.12%**	>	38.92%
rising	-5.67	52.70%	>	43.53%	**62.36%**	>	38.34%

Panel C - DJIA

Market phase	**Outperformance**	**Hit Rate**			**Abnormal Hit Rate**		
		long		**short**	**long**		**short**
falling	7.20	42.92%	<	58.77%	**53.37%**	>	45.97%
stagnant	-1.20	51.30%		47.99%	**54.98%**	>	44.94%
rising	-2.79	59.23%	>	47.37%	**59.76%**	>	35.06%

Notes: Panel A depicts the trading strategy performance measures (outperformance, hit rate, and abnormal hit rate) for both REITs and DJIA stocks. Panel B and Panel C show the three quality indicators for different market phases and trading results.

In order to find out how and in which market phase the trading strategy works most successfully for DJIA stocks, we apply the same methodology as for REITs and divide the time frame into falling, stagnant, and rising market phases and in buy or sell signals. The results for DJIA stocks (Panel C in Exhibit 2.4) show the same patterns as for REITs (Panel B). (1) The outperformance scores the highest value in falling market phases. (2) The hit rate for short (long) signals is higher in falling (rising) markets. (3) Regardless of the market phase, the abnormal hit rate for long signals is always substantially higher than for short signals. Therefore, we find that DJIA investors show the same information procurement and trading behavior, compared to REIT investors.

Nevertheless, outperformance, hit rate, and abnormal hit rate for REITs score better overall trading results especially in terms of predicting accuracy.

This raises the question as to why REITs score better than DJIA stocks. We hypothesize that the trading strategy with Google Trends data is better able to predict short-term movements of REITs, compared to DJIA stocks, because internet users supposedly leave more helpful traces online when searching for information about more homogeneous assets like REITs than for well-known, noisy DJIA stocks.

Due to the high popularity of the DJIA, the average number of investors for the DJIA is 1,760, compared to the relatively low average of 546 REIT investors.[6] Therefore, REITs supposedly capture a lower diversity of opinions. A higher number of investors with similar market projections subsequently leads to better trading results.

Furthermore, we assume that REIT investors tend to be more qualified and "pre-informed" than the average DJIA investor. Before googling for an explicit REIT, the potential investor already knows about the specifics of this particular investment vehicle and formulates his search queries accordingly. More precisely formulated search queries are less noisy and potentially generate more successful trading results.

Another characteristic that underpins better trading results towards REITs, is the homogeneity of the underlying asset. Every REIT has a high exposure to real estate. Therefore, the underlying asset is relatively homogeneous, compared to DJIA stocks, as they are assigned to nine different industries.[7] A higher heterogeneity of the underlying assets will increase sentiment dispersion and thus have a negative impact on the accuracy of stock price anticipation.

To summarize, two interesting empirical results arise from the detailed analysis of the trading strategy performance measures. First, we find that REITs perform better than DJIA stocks concerning outperformance, hit rate, and abnormal hit rate. Second, DJIA stocks show the same patterns in the context of different market phases and trading signals as REITs.

2.6 Conclusion

In our digitized society, the internet has evolved into the core information resource. Therefore, search volume data from Google Trends, which are freely available and easily accessible, provide valuable insights into our economic life on different levels.

Several researchers have already shown that there is a relationship between search volumes and stock returns and that search data has the potential to objectively reflect investors' underlying

[6] See: www.nasdaq.com as of January, 2017.

[7] The nine industries are: basic materials, consumer goods, consumer services, financials, health care, industrials, oil & gas, technology, telecommunications.

beliefs. Hence, information demand represented by Google search volumes is considered to be a good sentiment indicator (Da *et al.*, 2014). Search volume data has various advantages compared to conventional attention measurements. First, search volume data is freely available and, as Google is the search engine with the largest market share, it represents a significantly large sample at high frequency with only one hour delay. Second, search volume data captures attitudes rather than inquiries about them, and therefore has the potential to reveal personal interests more clearly than surveys. Finally, this kind of attention measure is likely to be an objective external measurement that verifies the actual information demand of a population (Da *et al.*, 2011; Beer *et al.*, 2013; Da *et al.*, 2014).

Although REIT prices are likely to be affected by sentiment just as general stocks (Lin *et al.*, 2009), the measurement of sentiment by search volume, its potential relationship with REIT prices, and its predicting abilities have not been examined comprehensively in the literature.

Empirically, we address three questions concerning the relationship between intraday search volumes and REIT prices. First, do Google search volumes have the potential to display investor sentiment and therefore have the ability to successfully predict REIT market movements? We incorporate a fictitious trading strategy for the Top 20 REITs of the MSCI US REIT Index, relying on relative changes of hourly search volumes. Following this approach, the performance of search volume-based trading strategies over 78 trading days is compared to simple buy-and-hold strategies/benchmarks of the underlying REITs. As for 75% of the REITs, the Google trading strategies outperform the benchmark, we conclude that company-related, hourly search volumes do certainly have the potential to predict intraday REIT price movements. On average, the trading strategy performance is 7.37 percentage points higher compared to a buy-and-hold strategy. The best performing REIT scores an outperformance of 45.60 percentage points, compared to the benchmark.

Second, in which market phases and before which trading signals do search volumes have the best prediction abilities for REITs? We divide the trading signals into long- and short-only and the period of consideration into falling, stagnant, and rising market phases. This differentiation allows us to obtain more detailed insights into the information-gathering behavior of internet users and as to whether information procurement plays the same role before buying as opposed to selling a stock. The outperformance shows that information demand has better predicting abilities in falling market phases, due to investors' concerns about the future. On average, the outperformance in falling markets is 14.06 percentage points higher than in rising markets. Furthermore, market participants tend to search more intensively for company-related information before buying a stock, rather than before selling it, as the hit rate for abnormal market movements is substantially higher for long signals, than for short signals.

Third, we investigate the much debated question of correlation and causality between search volumes and REIT prices. We find that, on average, there is a statistically significant negative correlation of -0.11, whereas individually, 25% of the REITs report a statistically significant positive correlation. The same phenomenon can be found for causality. On an aggregate level, the picture is very clear and we find a significant causality flow from prices to search volumes. But individually, we find bidirectional causality for 50% of the REITs. Furthermore, we validate our methodology by running the trading strategy for 30 DJIA stocks and show that REITs outperform the trading results of the DJIA Index.

As a whole, our findings suggest that hourly Google Trends search volumes with high granularity provide new and valuable insights into the search volumes' prediction abilities of REIT prices in the context of a hypothetical trading strategy. Previous research with weekly or monthly data has already shown that there is a relationship between online information demand and market movements in general. Incorporating detailed analyses of intraday search volume data enables us to determine the direction of this correlation and in which market phases the Google trading strategy has the best ability to predict future REIT market movements. This provides researchers on information demand and subsequent investment behavior with a solid base as to how information is processed or used to form investment decisions in REIT markets and, hence, how models and trading strategies should be set up.

2.7 Appendix

Appendix 2.1 | Descriptive statistics of search volumes

	Mean	**Median**	**Std. Dev.**	**Skewness**	**Kurtosis**	**Jarque-Bera**
Avalonbay Communities	70.96	73.00	14.76	-0.26	2.37	15.49
Boston Properties	73.47	76.00	14.50	-0.41	2.48	21.37
Digital Realty Trust	71.94	74.00	14.85	-0.34	2.40	18.69
Equinix	73.65	77.00	14.41	-0.52	2.66	27.79
Equity Residential	73.25	76.00	14.35	-0.47	2.61	24.01
Essex Property Trust	71.19	74.00	14.91	-0.28	2.33	17.80
Federal Realty Inv Trust	70.75	73.00	14.94	-0.25	2.31	16.78
General Growth Properties	71.22	74.00	14.59	-0.29	2.40	16.39
Welltower	69.25	70.00	14.43	0.01	2.38	8.81
HCP	76.58	79.00	13.28	-0.54	2.71	28.64
Host Hotels & Resorts	70.38	73.00	14.85	-0.24	2.35	15.36
Kimco Realty Corp	70.72	72.00	14.66	-0.24	2.32	15.92
Macerich	71.76	74.00	14.48	-0.31	2.45	15.76
Realty Income Corp	68.20	72.00	18.58	-1.26	5.86	335.47
ProLogis	72.63	75.00	14.89	-0.43	2.44	24.25
Public Storage	80.50	83.00	12.54	-0.86	3.49	73.28
SL Green Realty Corp	70.57	73.00	14.89	-0.24	2.35	14.78
Simon Property Group	71.87	74.00	14.61	-0.26	2.29	18.04
Vornado Realty Trust	70.16	73.00	15.25	-0.25	2.25	18.72
Ventas	64.87	64.00	16.33	0.01	1.95	25.54

2.8 References

Askitas, N. and Zimmermann, K.F. (2009), "Google Econometrics and Unemployment Forecasting", Applied Economics Quarterly, Vol. 55 No. 2, pp. 107–120.

Baker, M. and Wurgler, J. (2007), "Investor Sentiment in the Stock Market", *Journal of Economic Perspectives*, Vol. 21 No. 2, pp. 129–151.

Bank, M., Larch, M. and Peter, G. (2011), "Google search volume and its influence on liquidity and returns of German stocks", *Financial Markets and Portfolio Management*, Vol. 25 No. 3, pp. 239–264.

Barber, B.M. and Odean, T. (2008), "All That Glitters. The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors", *Review of Financial Studies*, Vol. 21 No. 2, pp. 785–818.

Barberis, N., Shleifer, A. and Vishny, R. (1998), "A model of investor sentiment", *Journal of Financial Economics*, Vol. 49 No. 3, pp. 307–343.

Barkham, R.J. and Ward, C.W.R. (1999), "Investor Sentiment and Noise Traders: Discount to net Asset Value in Listed Property Companies in the UK", *Journal of Real Estate Research*, Vol. 18 No. 2, pp. 291–312.

Beer, F., Hervé, F. and Zouaoui, M. (2013), "Is big brother watching us? Google, Investor Sentiment and the Stock Market", *Economics Bulletin*, Vol. 33 No. 1, pp. 454–466.

Black, F. (1986), "Noise", *The Journal of Finance*, Vol. 40 No. 3, pp. 529–543.

Canbaş, S. and Kandır, S.Y. (2009), "Investor Sentiment and Stock Returns. Evidence from Turkey", *Emerging Markets Finance and Trade*, Vol. 45 No. 4, pp. 36–52.

Chiang, K.C.H. and Lee, M.-L. (2010), "The Role of Correlated Trading in Setting REIT Prices", *The Journal of Real Estate Finance and Economics*, Vol. 41 No. 3, pp. 320–338.

Choi, H. and Varian, H. (2009), "Predicting the Present with Google Trends", *Working paper*, Google Inc., Mountain View, CA.

Clayton, J. and MacKinnon, G. (2001), "The Time-Varying Nature of the Link between REIT, Real Estate and Financial Asset Returns", *Journal of Real Estate Portfolio Management*, Vol. 7 No. 1, pp. 46–54.

Clayton, J. and MacKinnon, G. (2003), “The Relative Importance of Stock, Bond and Real Estate Factors in Explaining REIT Returns”, *The Journal of Real Estate Finance and Economics*, Vol. 27 No. 1, pp. 39–60.

Curme, C., Preis, T., Stanley, H.E. and Moat, H.S. (2014), “Quantifying the semantics of search behavior before stock market moves”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111 No. 32, pp. 11600–11605.

Da, Z., Engelberg, J. and Gao, P. (2011), “In Search of Attention”, *The Journal of Finance*, Vol. 66 No. 5, pp. 1461–1462.

Da, Z., Engelberg, J. and Gao, P. (2014), “The Sum of All FEARS Investor Sentiment and Asset Prices”, *Review of Financial Studies*, Vol. 28 No. 1, pp. 1–32.

Darrat, A.F., Rahman, S. and Zhong, M. (2003), “Intraday trading volume and return volatility of the DJIA stocks. A note”, *Journal of Banking & Finance*, Vol. 27 No. 10, pp. 2035–2043.

Das, P., Ziobrowski, A. and Coulson, N.E. (2015a), “Online Information Search, Market Fundamentals and Apartment Real Estate”, *The Journal of Real Estate Finance and Economics*, Vol. 51 No. 4, pp. 480–502.

Das, P.K., Freybote, J. and Marcato, G. (2015b), “An Investigation into Sentiment-Induced Institutional Trading Behavior and Asset Pricing in the REIT Market”, *The Journal of Real Estate Finance and Economics*, Vol. 51 No. 2, pp. 160–189.

DeLong, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. (1990), “Noise Trader Risk in Financial Markets”, *Journal of Political Economy*, Vol. 98 No. 4, pp. 703–738.

Dietzel, M., Braun, N. and Schäfers, W. (2014), “Sentiment-based commercial real estate forecasting with Google search volume data”, *Journal of Property Investment & Finance*, Vol. 32 No. 6, pp. 540–569.

Drake, M.S., Roulstone, D.T. and Thornock, J.R. (2011), “Investor Information Demand: Evidence from Google Searches Around Earnings Announcements”, *Journal of Accounting Research*, Vol. 50 No. 4, pp. 1001–1040.

Dzielinski, M. (2012), “Measuring economic uncertainty and its impact on the stock market”, *Finance Research Letters*, Vol. 9 No. 3, pp. 167–175.

Fama, E.F. (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work", *The Journal of Finance*, Vol. 25 No. 2, pp. 383–417.

García, D. (2013), "Sentiment during Recessions", *The Journal of Finance*, Vol. 68 No. 3, pp. 1267–1300.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7232, pp. 1012–1015.

Grullon, G., Kanatas, G. and Weston, J.P. (2004), "Advertising, Breadth of Ownership, and Liquidity", *The Review of Financial Studies*, Vol. 17 No. 2, pp. 439–461.

Guzman, G. (2011), "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations", *The Journal of Economic and Social Measurement*, Vol. 36 No. 3, pp. 119–167.

Hohenstatt, R. and Käsbauer, M. (2014), "GECO's Weather Forecast for the U.K. Housing Market: To what Extent can we rely on Google Econometrics?", *Journal of Real Estate Research*, Vol. 36 No. 2, pp. 253–281.

Hohenstatt, R., Käsbauer, M. and Schäfers, W. (2011), ""Geco" and its Potential for Real Estate Research: Evidence from the U.S. Housing Market", *Journal of Real Estate Research*, Vol. 33 No. 4, pp. 471–506.

Hu, W.-Y., Huang, C.-J., Chang, H.-Y. and Lin, W.-J. (2015), "The Effect of Investor Sentiment on Feedback Trading and Trading Frequency. Evidence from Taiwan Intraday Data", *Emerging Markets Finance and Trade*, Vol. 51 No. 1, pp. 111–120.

Joseph, K., Wintoki, M.B. and Zhang, Z. (2011), "Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment Evidence from Online Search", *International Journal of Forecasting*, Vol. 27 No. 4, pp. 1116–1127.

Kahneman, D. and Tversky, A. (1979), "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, Vol. 47 No. 2, pp. 263–291.

Lin, C.Y., Rahman, H. and Yung, K. (2009), "Investor Sentiment and REIT Returns", *The Journal of Real Estate Finance and Economics*, Vol. 39 No. 4, pp. 450–471.

McLaren, N. and Shanbhogue, R. (2011), "Using internet search data as economic indicators", *Bank of England Quarterly Bulletin*, Q2, pp. 134–140.

Preis, T., Moat, H.S. and Stanley, H.E. (2013), "Quantifying trading behavior in financial markets using Google Trends", *Scientific reports*, Vol. 3 No. 1684, pp. 1–6.

Preis, T., Reith, D. and Stanley, H.E. (2010), "Complex dynamics of our economic life on different scales: insights from search engine query data", *Philosophical Transactions of the Royal Society*, Vol. 368 No. 1933, pp. 5707–5719.

Ro, S. and Gallimore, P. (2014), "Real Estate Mutual Funds. Herding, Momentum Trading and Performance", *Real Estate Economics*, Vol. 42 No. 1, pp. 190–222.

Rochdi, K. and Dietzel, M. (2015), "Outperforming the benchmark: online information demand and REIT market performance", *Journal of Property Investment & Finance*, Vol. 33 No. 2, pp. 169–195.

Rosenberg, B., Reid, K. and Lanstein, R. (1985), "Persuasive evidence of market inefficiency", *The Journal of Portfolio Management*, Vol. 11 No. 3, pp. 9–16.

Vlastakis, N. and Markellos, R.N. (2012), "Information demand and stock market volatility", *Journal of Banking & Finance*, Vol. 36 No. 6, pp. 1808–1821.

Wu, L. and Brynjolfsson, E. (2009), "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales", *ICIS 2009 Proceedings*, pp. 147–171.

3 Leveling the Playing Field: Out-of-town Buyer Premiums in US Housing Markets Over Time

Abstract

Purpose – Historically, research shows that out-of-town buyers of real estate are informationally disadvantaged and therefore pay higher prices compared to in-town buyers. However, with the recent advent of online housing platforms, a plethora of information about the housing market is provided for free. The purpose of this paper is to examine whether out-of-town buyers do in fact pay a premium and why, and whether this premium has decreased due to better information availability.

Design/Methodology/Approach – A hedonic regression model over a ten-year window (2005, 2015) is developed to analyze condominium transactions in Miami-Dade County. The results are validated by various robustness checks and the propensity score matching algorithm to identify a comparable control sample for 2015 in terms of relevant housing characteristics.

Findings – The results support the hypothesis that out-of-town buyers pay higher prices for real estate than their local counterparts, and that both search costs and anchoring cause a premium in both years whereas wealth only plays a significant role in 2005. The premium due to search costs and therefore information availability has decreased slightly over time.

Originality/Value – This is the first out-of-town paper that compares two points in time versus a single cross section analysis. Besides the premium caused by information asymmetry/search costs measured by distance and the anchoring effect, the regression model is extended by the wealth effect.

3.1 Introduction

Real estate markets are characterized by heterogeneity, infrequent trading, geographic segmentation, and information asymmetry. One of the most important differences between a perfectly competitive market and the housing market is the heterogeneity in the cost of acquiring information, due to information asymmetry among different buyer types. If the level of information or search costs vary systematically with buyer characteristics, then the price paid for any property should also vary with buyer characteristics.

In particular, past research has found that out-of-town buyers "overpay" for real estate, compared to local buyers. The reason why out-of-town buyers pay more, is shown to be twofold. Firstly, local buyers face lower search costs, are better informed, and have more bargaining power than out-of-town buyers, and therefore, pay lower prices compared to out-of-town buyers (Miller and Rice, 1981; Lambson *et al.*, 2004; Clauretie and Thistle, 2007). Secondly, researchers have shown that out-of-town buyers from areas where real estate prices are high relative to local prices may have upwardly biased expectations of property values and therefore overpay (Ihlanfeldt and Mayock, 2012; Zhou *et al.*, 2015). This phenomenon results from a cognitive bias and has been termed the "anchoring effect" (Tversky and Kahneman, 1974). Ling *et al.* (2016) confirmed both reasons, namely that out-of-town buyers pay a premium for real estate due to informational disadvantages and anchoring. They stated that these pricing outcomes are more likely to be observed in segmented and informationally inefficient markets.

Although the effects of different search costs in relation to geographical distance and behavioral bias on pricing outcomes have been documented in the literature, the possible effects of differences in information availability in the last few years have not yet been addressed.

Therefore, we aim to answer three research questions in this paper: (1) Do out-of-town buyers pay more for real estate compared to locals? (2) If so, is that premium caused by different search costs (distance), biased beliefs (anchoring), or the level of personal income (wealth)? (3) Did the search cost premium decrease from 2005 to 2015 due to better public information?

A similar question has already been addressed by Turnbull and Sirmans (1993). They investigated whether existing institutions, such as the multiple listing service (MLS) and mortgage lending requirements, provide sufficient information to protect less-informed buyers from systematically purchasing houses for more than they are worth. Of course, several subsequent studies have suggested that mortgage lenders do not protect buyers from overpaying which helped to fuel the housing crisis of 2007-2008 in the US.[8] Here, we pursue the original question of Turnbull and Sirmans (1993), but focus on the search cost and information-availability side. We aim to find out whether the enormous increase in information availability by the internet over the last ten years has assisted previously informationally disadvantaged out-of-town buyers by reducing the probability of overpaying.

The reason why we believe it is important to investigate especially the last question is that there has been an enormous rise in information availability over the last decade. With the launch of home search websites like Zillow, Trulia, or Redfin since 2006, a plethora of information about the housing market is now provided for free. These websites act as real estate and home-related information marketplaces, empowering homeowners, buyers, sellers, renters, professionals, and lenders of all types with data, inspiration, and various forms of knowledge about the real estate market. Zillow's database, for example, covers more than 110 million US homes, and "Zestimate" home values, prior purchase prices, rents, and other home-related information (Zillow, 2017).

By acting as electronic marketplaces, these home search websites tend to equalize the information level between heterogeneous buyers and sellers. If information asymmetry or search costs decrease systematically with an increased availability of information, then the prices paid for comparable properties should be more equal, exhibiting less price dispersion and a lower price premium, and that for all buyer types (Byrnjolfsson and Smith, 2000).

We therefore examine whether different categories of homebuyers (out-of-town and in-town buyers) pay significantly different prices for comparable houses, and whether this difference has changed between 2005 and 2015. These two years are chosen, because they constitute a sufficient time gap associated with an enormous improvement in information availability.

This paper makes a research contribution by extending previous research in the following ways. First, we examine whether out-of-town buyers continue to pay a premium for real estate over a ten year window in Miami-Dade County, versus a single cross section in time. By doing so, we can compare two points in time with differences in terms of information availability. Second, we concentrate on non-owner-occupied condominium transactions, a previously neglected segment of the housing market. Third, we wish to find out whether this premium is caused by information asymmetry/search costs measured by distance or the anchoring effect, and we extend the model by the wealth effect, as we assume that this independent effect was measured as part of the anchoring effect in previous studies. Fourth, by applying the propensity score matching algorithm,

[8] Recent academic literature confirms that appraisal bias and inflation were pervasive among loans originating and sold into RMBS during the 2002-2007 period. According to one of these studies, Griffin and Maturana (2016) examined over three million loans that were sold into non-agency RMBS between 2002 and 2007. With the aid of a retrospective AVM, the authors found that the appraisals were overstated (by 20% or more) for 13.2% of purchases and 20.5% of refinances, an average of 17.8% of the loans overall. Similar results have been published by Agarwal *et al.* (2015).

we bypass the selection bias and identify a comparable control sample of transactions in 2015 in terms of the relevant housing characteristics. The regression result of the matched sample serves as a robustness check and strongly supports our results for all three variables of interest (distance, anchoring, and wealth). Finally, we look not only at one city or county, but make an intercounty comparison in order to test the generalizability of the results.

By incorporating both a theoretical search model and a hedonic pricing regression, this study investigates whether or not heterogeneous buyers are more equally informed nowadays than before the internet revolution. This question is part of a broader set of concerns over the impact of asymmetric information and search costs in the pricing system and is therefore relevant to the broader literature.

Our results show that distant buyers continue to pay a price premium, relative to local buyers in Miami-Dade County in 2015, although the premium in 2015 is lower than in 2005. This decrease may be due to better public information, despite the standard deviation of both sales prices and prices per square foot rising from 2005 to 2015. Another interpretation of the result is that local agents continue to exploit asymmetric information gaps and/or higher search costs, which results in a continuation of the premium. By decomposing the sources of these premiums, we find that search costs associated with buyer distance are both economically and statistically significant in explaining observed price premiums. Behavioral bias in the form of anchoring tends to play a less important role, as the premium is statistically significant but very small. Macroeconomic factors such as personal income (wealth effect) are only statistically significant in 2005. Finally, we validate our results with a propensity score matching and various robustness checks and find an important caveat. The results can be confirmed for the same region, but cannot be generalized for other counties in the US, as San Francisco County and San Diego County do not produce the same findings as Miami-Dade.

The rest of the paper is structured as follows. The second section surveys the relevant literature and the theoretical background in terms of a search model. The third section describes our empirical model, while the fourth contains the data description and summary statistics. Our regression results and robustness checks are presented in the fifth section, together with an intercounty comparison. Conclusions are drawn in the sixth section.

3.2 Literature and Theoretical Background

3.2.1 Literature Review

Different groups of real estate buyers face different search costs and tend to have heterogeneous information levels. For example, locals know more about location-related information such as trends in growth, zoning, crime, and so forth. Given these deviations from a perfect market, a divergence from the law of one price seems plausible (Lambson *et al.*, 2004; Horenstein *et al.*, 2017).

Some research has been undertaken about information asymmetry, search costs, and anchoring and their impact on house prices, using a variety of definitions for out-of-town buyers and for anchoring.

The first research strand focuses on out-of-town buyer premiums caused by information inefficiency and search costs. Miller *et al.* (1988) concentrate on Japanese buyers in two Honolulu areas in the late 1980s and show that Japanese buyers paid significantly more than local buyers. Ten years later, Watkins (1998) indicates that distant entrants to a local residential property market in Glasgow, Scotland, do not pay significantly different prices for a hypothetical standard property. As mentioned, Turnbull and Sirmans (1993) find a positive but insignificant premium for 151 single-family houses in Baton Rouge, Louisiana, from May 1988 to June 1989. This study was revisited by Ihlanfeldt and Mayock (2012), who strongly support the distant buyer hypothesis when using distance to identify non-local buyers of single-family houses in Florida. Horenstein *et al.* (2017) examine 1,013 repeat sales of small ranches used for agriculture and recreation from 1991-2000 to show that local buyers obtain greater returns than out-of-town buyers when they resell their properties.

The second research strand focuses explicitly on the anchoring effect. Tversky and Kahneman (1974) were the first to write about anchoring as a cognitive bias. They suggest that individuals use arbitrary reference values or anchors that influence their estimate of value, and that they only deviate marginally from these reference values (anchors). With regard to the real estate market, this implies that buyers use information about their own familiar real estate market, rather than undertaking research about another unknown real estate market, to make initial estimates about mean house prices. The first real estate application of the anchoring effect was conducted by Northcraft and Neale (1987), who study the anchoring bias in amateur and expert valuations of real estate. Ten years later, Diaz and Hansz (1997) show that appraisers valuing properties in unfamiliar geographic areas take the opinion of an "anonymous expert" as an anchor, thus affecting their valuations.

The third research strand constitutes a combination of both effects (search costs and anchoring). With Zhou *et al.* (2015) and Ling *et al.* (2016), two more recently published papers support both the distant and anchoring hypotheses. Zhou *et al.* (2015) use 940 transaction data from a large development in Chengdu, China, from 2009-2011, whereas Ling *et al.* (2016) use 18,372 industrial, multi-family, and office-sale transactions of the fifteen largest MSAs in the US. Lambson *et al.* (2004) investigate 2,854 apartment transactions in the Phoenix area and find that out-of-state buyers pay a statistically significant premium compared to in-state buyers.

Furthermore, they find weak evidence that buyers from high-price states pay more than buyers from non-high-price states, which can be attributed to the anchoring effect. We hypothesize that this phenomenon can also be provoked by higher personal income rather than anchoring. In order to test whether this holds true for our model we include the wealth variable.

The most important empirical studies about the out-of-town buyer premium and the anchoring effect are concisely shown in a table in Appendix 3.1.

In summary, previous research has focused mainly on prices paid for single-family or commercial properties for one period of time, not taking into account the possible changes in information availability and therefore search costs in the past few years by comparing two points in time.

3.2.2 Search Model

By incorporating a sequential-search model, we identify potential price effects with respect to differential search costs and anchoring behavior. The search model can be described mathematically as an optimal search-stopping problem and builds on prior real estate search models such as Miller and Rice (1981), Turnbull and Sirmans (1993), Lambson *et al.* (2004), and Zhou *et al.* (2015). In a first step, we introduce the common assumptions of our search model followed by the scenarios of changing search costs and the anchoring effect.

Model Assumptions

Market participants can be separated into two groups: Firstly, there are sellers, who are willing to sell their house at different reservation prices, where $F(P)$ is the distribution of prices per condominium with the sales price $P \in \{0,\infty\}$, which is independent of offers in prior rounds. Secondly, there are buyers, who are differentiated in terms of market information level. That is, some buyers are aware of more details among alternatives that might suit their housing demands better. Thus, we distinguish between more informed and less informed buyers. For simplicity reasons, we assume risk-neutral buyers and a zero discount rate.

All buyers face constant marginal search costs $SC \in \{0,\infty\}$, which vary across buyers and are associated with finding an appropriate property to purchase. We assume that a buyer who is at an informational disadvantage incurs higher marginal search costs than a more informed buyer. The search costs SC include, for example, the costs of travel, physical and financial inspection, studies of the neighborhood, the local economy and demographics, historical research on the property including improvements, past prices, defects, and negative externalities surrounding the property (Lambson *et al.*, 2004). All search costs include an opportunity cost of time aside from direct costs.

The most important assumption in our search model is that rational investors continue searching for better value until the marginal search cost of an additional search equals the marginal benefit or expected value of the next search (Stigler, 1961). Hence, the optimal search is a trade-off between getting a lower price by searching one more time, and the cost of continued searching. This assumption implies that investors with high search costs will search less than investors with low search costs and consequently overpay on average, because the stopping rule will trigger earlier (Miller and Rice, 1981).

The buyer anticipates an exogenous value $V \in \{0,\infty\}$, from the particular property under consideration. We assume that the aim of every buyer is to maximize the net value or surplus $V - P$ from purchasing at price P.

To include anchoring, we add the variable $A \in \{-\infty,\infty\}$ which is a parameter that shifts some of the probability weight, based on a priori distributions in the buyer's mind, prior to starting any search process. Buyers enter the market with different beliefs about the price distribution *F(P,A)*, which is the probability that the buyer will find an offer at $P \in \{0,\infty\}$, which is less than or equal to the optimal reservation price P^*, and A. If A is positive, for example, it shifts the probability weight from lower to higher prices. Thus, it indicates that some buyers perceive the distribution of prices as higher than they actually are, due to anchoring. Note that we could also have the converse, where buyers from less expensive markets assume that the distribution of available properties should be lower than they actually are.

Again, and in an integrated model, the decision to stop searching is a decision about an optimal reservation price $P^* \in \{0,\infty\}$. Only offers below the reservation price ($P \leq P^*$) are accepted.

The stopping rule P^* should satisfy[9]:

$$(V - P^*) = [1 - F(P^*, A)](V - P^*) + \int_0^{P^*} (V - P)F(P, A)dP - SC \tag{5}$$

By integrating up to the optimal reservation price P^*, the expected value of finding an appropriate price results. Rearranging (5) yields:

$$SC = \int_0^{P^*} (P^* - P)F(P, A)dP \tag{6}$$

The optimal reservation price is reached when the buyer is indifferent between stopping and continuing the search. That means that the marginal cost of searching (left-hand side of equation) is equal to the expected marginal benefit of an additional search (right-hand side of equation), which depends on the probability of securing a price lower than P^*.

[9] The variable description and parts of equation (5) are shown in Appendix 3.2.

Differing search costs: the case of out-of-town buyers

As assumed, buyers who are informationally disadvantaged face higher marginal search costs than informed buyers. To identify the impact of differing levels of search cost on the optimal reservation price $P*$, we rearrange (6) and differentiate with respect to *SC*:

$$\frac{\partial P^*}{\partial\, SC} = \frac{1}{F(P^*, A)} \quad > \ 0 \tag{7}$$

With an increase in search cost *SC*, the optimal reservation price $P*$ will increase as well. Intuitively, when the search costs and the reservation price for one group of buyers is higher, this group of buyers will, on average, stop searching earlier and pay higher prices than their counterparts with lower search costs.[10]

In our case, we assume that out-of-town buyers are informationally disadvantaged and therefore suffer higher search costs. Due to higher search costs, distant buyers are expected to pay higher prices, on average, compared to local buyers. Again, the reason we suggest that out-of-town buyers have a lower information level is because locals have lower search costs. Prior to starting a search, they already know more about the local market. They may hear and read local news. They absorb information from driving, working, and shopping in the area. The time and money spent gathering information with respect to physical and financial information is quite limited, compared to out-of-town buyers (Lambson *et al.*, 2004). With our search model, we show theoretically that with presumably higher search costs for out-of-towners, the prices are consequently higher compared to locals.

Differing price perception: the case of anchoring

Buyers from different cities may have different perceptions or beliefs about the distribution of prices *F(P,A)* with $A \in (-\infty,\infty)$, which shifts the probability weight. To identify how the optimal reservation price $P*$ changes due to changes in A, we rearrange and differentiate equation (6) with respect to A:

$$\frac{\partial P^*}{\partial\, A} = \frac{-\int_0^{P^*} (P^* - P) f_A dP}{F(P^*, A)} \quad > \ 0 \tag{8}$$

Buyers may expect the prices to be higher than the underlying price distribution really is, due to anchoring. With these beliefs, these buyers will accordingly set their optimal reservation price $P*$, on average, higher and settle for higher prices than buyers without anchors. Intuitively, they will stop their search sooner, as $P*$ and P will be higher. We only address the case of positive anchoring here, but the reverse could also apply, where $A < 0$ and this affects the optimal

---

10 Caveat: (7) does not automatically imply that all buyers with higher search costs will pay a premium. Some, for example, could get a price *P*, which is lower than $P*$ on the first search by chance.

reservation prices a buyer is willing to pay. Intuitively, a higher reservation price might lead to stopping sooner than otherwise, based on the lower probability of benefits from continued searching.

3.3 Empirical Methodology

To estimate the effect of search costs, anchoring, and wealth on the transaction price, we estimate the following multivariate stepwise hedonic regression model:

Model I $\ln(PRICE)_i = \beta_0 + \beta_1 SQFT + \beta_2 NBR\ BEDRMS + \beta_3 NBR\ BATH + \beta_4 \ln(AGE) + \beta_5 LOCATION + \beta_6 INST\ INV + \beta_7 GDP + \varepsilon_i$

Model II $+\beta_8 \ln(DISTANCE) + \varepsilon_i$

Model III $+\beta_8 \ln(DISTANCE) + \beta_9 ANCHOR + \varepsilon_i$

Model IV $+\beta_8 \ln(DISTANCE) + \beta_9 ANCHOR + \beta_{10} WEALTH + \varepsilon_i$ (9)

Dependent variable

ln(PRICE)	Is the natural logarithm of the selling price for a condominium. The natural logarithm transformation enables normalizing the strong left-skewed distribution of the selling price.

Property and transaction variables

SQFT	Is the number of square feet of living space.
NBR BEDRMS	Is the condominium's number of bedrooms.
NBR BATH	Is the condominium's number of bathrooms.
ln(AGE)	Is the natural logarithm of the condominium's age in years. We include the transformation in order to consider the non-linear relationship between age and selling price.
LOCATION	Is the time, measured in minutes that it takes to drive from the condominium to the city center of Miami. This variable yields more information about the location of the condominium within the local community. It is calculated from the longitude and latitude coordinates of the transacted condominium and the city center of Miami, by using R in combination with the OpenStreetMap-Based Routing Service (Kohle and Wickham, 2013; Giraud *et al.*, 2017).
INST INV	Is a category variable, which is set equal one if the transaction was carried out by an institutional investor, otherwise, it is set equal zero.
GDP	Is the GDP of the years 2005/2015 in million dollars on a MSA level.

Out-of-town variable

ln(DISTANCE)	Is a simple straight-line distance in miles between the condominium and the buyer's listed address. It is calculated from the longitude and latitude coordinates. We use the natural logarithm of distance as an out-of-town variable, because there is no linear relationship between distance and price. Previous literature has shown that distance is an appropriate measure for search costs (Smith *et al.*, 1999; Clauretie and Thistle, 2007; Ihlanfeldt and Mayock, 2012).

Anchoring variable

ANCHOR	Is the median price per square foot in dollars for condominiums in 2005/2015 of the buyer's origin according to the zip code. This variable replicates the buyer's beliefs about prices, which may be upwardly or downwardly biased.

Wealth variable

WEALTH	Is the average personal income per capita in 2005/2015 in dollars of the buyer's origin minus the average personal income per capita of the property's site on a MSA level.[11] This variable is a good proxy for indicating whether the buyer is from a rather more or less wealthy area. Therefore, if the buyer is from a wealthier region compared to the property's region, the variable is positive. If the buyer is not, it is negative. We include this variable in order to investigate whether distant buyers pay more, due to a higher income level or rather due to higher search costs measured in distance and biased (incorrect) beliefs.

The regression model comprises four models which build on one another. The empirical baseline *Model I* is a simple multivariate stepwise regression, with the price for a given property as a function of its characteristics. The last three variables (distance, anchor and wealth) are not taken into account in the baseline model. The out-of-town *Model II* extends the baseline model for the natural logarithm of distance as an out-of-town variable, which allows us to determine whether the prices increase with the distance of buyers in miles. We use distance as an out-of-town variable for two reasons. Firstly, it is a continuous variable and possesses more explanatory power than a binary out-of-town indicator. Secondly, research has shown that distance is a good proxy for search costs and therefore information availability, as the two are regarded as being positively correlated (Zhang *et al.*, 2016). As there is no direct measure available that accounts for the changes in information levels between 2005 and 2015, we use distance as a proxy for information availability. Hence, if the premium caused by distance reduces from 2005 to 2015, this translates to a reduction of search costs over time. This reduction is presumably related to better information levels among buyers from 2005 to 2015.

The anchoring *Model III* includes both the out-of-town indicator and the anchor, in order to determine which effect has a greater impact on the selling price. The last *Model IV*, which is our

[11] The data is available under: https://www.bea.gov/.

overall model as shown in equation (9), includes the out-of-town indicator (distance), anchoring and wealth. With this model, we aim to show which effect leads to the highest premium.

Quadratic transformations of all continuous explanatory variables are employed by Weirick and Ingram (1990), Lambson *et al.* (2004), and Clauretie and Thistle (2007) in the regression models. We do not adopt their approach, due to multicollinearity issues. In order to quantify the severity of multicollinearity, we calculate a correlation matrix including all variables used in the regression, which can be found in Appendix 3.3. The estimated correlation between anchor and the selling price is 0.22. Between wealth and the selling price it is 0.12. As expected, the highest correlations (0.72 to 0.75) arise between number of bedrooms, number of bathrooms, and living space in square feet. Nevertheless, there is no serious issue of multicollinearity.

3.4 Data

3.4.1 Sample Composition

The data used in this paper include condominium transactions that occurred in Miami-Dade County in 2005 and 2015. The Public Records data was provided by Collateral Analytics. Our overall dataset consists of 15,795 transactions from which 5,446 took place in 2005 and 10,349 in 2015. The original sample of 44,385 transactions was reduced in the course of the following data-cleaning steps.

Firstly, we concentrate on non-owner-occupied condominiums only, because owner-occupied properties do not provide information on where the buyer originally came from. This approach is similar to Clauretie and Thistle (2007). Secondly, transaction records missing one or more of the variables required for our hedonic regressions are excluded. Thirdly, we took a closer look at the distribution of sales prices and prices per square foot and include only transactions from p5- to p95-quantile, in order to eliminate outliers. Finally, all transactions with more than four bedrooms and more than six bathrooms are excluded, as well as transactions with the combination of zero bed- and zero bathrooms. The data is cleaned, in order to eliminate, early on, atypical properties and to ensure a representative sample which is as homogeneous as possible.

For homogeneity reasons, we also restrict our sample to condominiums only, following Ihlanfeldt and Mayock (2012) and Zhou *et al.* (2015), even if the market for non-owner occupied condominiums behaves differently from the broader housing market. A significant part of price dispersion and biased regression estimates can evidently be attributed to the heterogeneous nature of real estate assets. Whereas a homogeneous sample with regard to property type and housing characteristics, more effectively isolates the effects of asymmetric buyer information, anchoring, and wealth, thus eliminating much of the noise, and reducing the potential of omitted variables. The reason why we focus on Miami-Dade County is that the real estate market there contains a sufficient number of non-owner occupied condominiums that are bought by out-of-town buyers. The availability and quality of the data has shown to be another reason for choosing this county.

With our paper, we aim to show if and how increased information availability affects the out-of-town buyer premium. A ten year window between 2005 and 2015 is chosen, in order to replicate a sufficient time gap to meaningfully demonstrate the impact of increased information availability. In 2006, Zillow, one of the leading real estate data providers in the US, was founded. Afterwards, with the launch of several other home search websites and real estate data providers like Redfin, Trulia, Realtor, Homes, etc. in the past few years, a veritable plethora of information about real estate prices, home and location characteristics, past sales price histories for homes, and even value-forecasting for homes, are provided for free. Zillow, for example, claims to have "Zestimates" for more than 100 million homes, with 100-plus attributes tracked for each property. However, researchers show that Zillow's estimates of home values, so called "Zestimates", are overpriced by 11.66% on average (Hollas *et al.*, 2010). But even if Zillow overestimated the value for 40% of the properties by more than 10%, the "Zestimate" value functions as a reference point and trend indicator which was not available in 2005.

With our time scope 2005 and 2015, we wish to depict one year (2005) in which real estate websites barely existed and all the types of information named above was only accessible with considerable effort and search costs. The second year (2015) is intended to represent one in which information availability changed enormously, and information about real estate prices is provided for free on a number of well-established home search websites. To give an overview of the difference in information availability, a table in Appendix 3.4 shows a comparison of information levels between 2005 and 2015.

Critics may argue that information about real estate markets has already been available over the internet in 2005 and information availability did not increase enormously over time. In theory, it was available online from county records since pre-2005, but quite limited in scope and requiring numerous searches to aggregate any market range data with no history of prices. The effort and time to receive information has changed tremendously from 2005 and 2015, as home search websites act as electronic marketplaces which bundle information from different data providers. Moreover, the internet increasingly offers applications such as Google Street View or virtual 360 degree videos of homes. These innovations reduce laborious, time-intensive research and therefore help to equalize the information level between heterogeneous buyers and sellers.

With respect to the sample periods, both 2005 and 2015 were up-cycle phases. For better comparability, it is important to replicate two years in roughly the same phase of a real estate cycle. Therefore, several macroeconomic factors like employment, GDP, earnings, population, and personal income from 2005 and 2015 are compared. The results show that the real changes in these factors between the two years yield not more than 10%. The number of observations, however, does not support the impression of same cycle phases, as it is twice as high in 2015 as in 2005. This phenomenon is caused by data availability in 2015 (48% more transactions), and the fact that the non-owner-occupied rate increased from 36% to 54% in these two samples.

3.4.2 Descriptive Statistics

Exhibit 3.1 reports the descriptive statistics for the whole sample (Panel A) and for the out-of-town (Panel B) and in-town buyer (Panel C) sub-samples. Considering the overall sample, the mean sales price for a condominium in Miami-Dade was $275,837 in 2005. The typical transaction in 2005 is a 1,098 square feet condominium with approximately 2 bedrooms and 1.5 bathrooms. A mean building age of 21 suggests that the properties are relatively new, especially compared with northern cities. The average buyer lives 324 miles away from the property and the median price per square foot for a condominium of the buyer's origin is $245. For 2015, the mean sales price was $281,601 and the difference between 2005 and 2015 is not statistically significant, with a *t*-statistic of -1.36. Further property characteristics in 2015 were pretty similar to 2005, except for the mean property age (29 years) and the institutional investor participation. The distance of the buyer reduced to 201 miles and the difference between the income levels (wealth) increased from $684 in 2005 to $1,114 in 2015. The statistics for the out-of-town (Panel B) and in-town buyer (Panel C) sub-samples clearly show differences for the variables sales price, square feet, institutional investor participation, GDP, distance, anchor, and wealth.

This paper aims to show how and if the out-of-town buyer premium cause price changes between 2005 and 2015. Intuitively, if search costs decrease systematically with increased information availability, then the prices paid for any properties should be more equal, exhibiting less price dispersion for all buyer types (Rothschild, 1974; Pereira, 2005). Hence, one can assume that the standard deviation of prices in 2015 is lower than in 2005, due to better public information. The empirical "anomaly" known as price dispersion is probably one of the most important distinctive features of housing markets (Maury and Tripier, 2014). The literature has mainly responded to the price dispersion phenomenon by introducing the heterogeneity of buyers and sellers, especially in the light of search costs (Lisi, 2014).

Exhibit 3.1 | Descriptive statistics

Panel A	2005		2015		
All properties (15,795 obs.)	**5,280 obs. (100%)**		**10,515 obs. (100%)**		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	*t*-statistic
Sales price ($)	275,837	214,279	281,601	267,364	-1.36
Square feet	1,098	454	1,109	445	-1.55
Price psqft ($)	249	121	248	170	0.57
Number of bedrooms	1.83	0.83	1.88	0.87	-3.93
Number of bathrooms	1.66	0.63	1.73	0.68	-5.58
Age (years)	20.65	16.95	28.56	19.45	-25.16
Driving time to center (minutes)	22.56	10.37	22.05	11.40	2.72
Institutional Investor	0.12	0.32	0.58	0.49	-61.71
GDP ($)	344,402	299,523	388,516	310,450	-8.36
Distance (miles)	324	563	201	470	13.26
Anchor ($)	245	149	259	213	5.28
Wealth ($)	684	3,817	1,114	5,237	-5.21

Panel B					
Out-of-town buyers (5,446 obs.)	**2,150 obs. (41%)**		**3,296 obs. (31%)**		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	*t*-statistic
Sales price	337,537	252,924	338,516	301,380	-0.12
Square feet	1,162	511	1,148	485	1.05
Price psqft	287	131	288	176	-0.15
Number of bedrooms	1.82	0.84	1.83	0.88	-0.26
Number of bathrooms	1.70	0.64	1.74	0.71	-1.66
Age (years)	20.90	18.35	30.77	20.38	-18.17
Driving time to center (minutes)	22.00	9.82	21.27	10.08	2.64
Institutional Investor	0.09	0.29	0.42	0.49	-27.72
GDP ($)	504,629	437,065	559,885	537,742	-3.79
Distance (miles)	1,053	542	1,085	536	-1.61
Anchor ($)	305	221	332	357	-1.10
Wealth ($)	1773	5,990	3,821	9,150	-8.74

Panel C					
In-town buyers (10,349 obs.)	**3,130 obs. (59%)**		**7,219 obs. (69%)**		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	*t*-statistic
Sales price	233,455	170,638	255,616	245,983	-4.58
Square feet	1,054	406	1,092	425	-4.29
Price psqft	224	107	230	164	-1.93
Number of bedrooms	1.84	0.83	1.91	0.87	-4.26
Number of bathrooms	1.63	0.62	1.72	0.67	-6.18
Age (years)	20.47	15.93	27.55	18.92	-18.31
Driving time to center (minutes)	22.95	10.71	22.41	11.95	2.17
Institutional Investor	0.14	0.35	0.66	0.47	-55.02
GDP ($)	243,906	1,327	317,986	0	-4700
Distance (miles)	10	8	10	8	0.11
Anchor ($)	215	75	233	116	4.29
Wealth ($)	0	0	0	0	-

Notes: Exhibit 3.1 presents the descriptive statistics of all relevant variables used in the baseline regression for the years 2005 and 2015. The overall sample can be divided into out-of-town and in-town buyer sub-samples. A out-of-town buyer is defined as a buyer who lives more than 50 miles away from the property he purchases. The t-statistic reports whether the difference between the years is statistically significant.

However, as shown in Exhibit 3.1, the price dispersion (Std. Dev.) for sales prices, as well as for prices per square foot, increases from 2005 to 2015, despite a higher level of information availability for the whole sample and for the subcategories. The standard deviation for the sales price for out-of-town buyers in 2005 was $252,924 and it increased to $301,380 in 2015 (Panel B). These results may seem contrary to the likely impact of better information efficiency. A few empirical studies found the same paradox. Byrnjolfsson and Smith (2000), Biswas (2004), and Pereira (2005) stated that higher price dispersion in the context of better information availability can arise from a higher level of product and information differentiation. Especially in heterogeneous markets, reduced search costs caused by the internet can in fact lead to greater price dispersion.

Our first research question addresses whether out-of-town buyers pay more for real estate than local buyers. Therefore, we take a closer look at the sales prices, square feet, and prices per square foot by out-of-town buyer and in-town buyer for both years in Exhibit 3.2.

Exhibit 3.2 | Statistics: out-of-town buyers vs. in-town buyers

	2005				2015			
All properties	OOTB Mean	ITB Mean	Difference	*t*-statistic	OOTB Mean	ITB Mean	Difference	*t*-statistic
Sales price	337,538	233,455	-104,082***	-17.86	338,516	255,616	-82,900***	-14.90
Square feet	1,162	1,054	-108***	-8.58	1,148	1,092	-56***	-5.93
Price psqft	287	224	-63***	-19.25	288	230	-58***	-16.36

Notes: OOTB Mean is the mean selling price paid by out-of-town buyers and ITB Mean is the mean selling price paid by in-town buyers in $. The difference is calculated by ITB Mean – OOTB Mean. * Indicates significance at the 10% level. ** Indicates significance at the 5% level. *** Indicates significance at the 1% level.

In 2005, out-of-town buyers paid $337,538 on average for a condominium, compared to in-town buyers with $233,455. Hence, they pay about 45% more on average in 2005 and about 32% more in 2015 than in-town buyers. They tend to purchase condominiums with more square feet of living space than in-town buyers in both years. The prices per square foot for out-of-town buyers were 28% higher in 2005 and 25% higher in 2015 compared to local buyers. Thus, the premium paid by out-of-town buyers declined significantly from 2015 to 2005. Given the sample size, all the differences in the means between out-of-town and local buyers are statistically significant for both years. These statistical significant differences, however, do not constitute evidence that out-of-town buyers really pay too much for condominiums in Miami-Dade County. Rather, this is only evidence that non-locals bought expensive properties. Determining whether they overpay requires controlling for covariates such as housing characteristics in our hedonic regression.

3.5 Results

3.5.1 Regression Results

Model I: Our baseline model is a standard multivariate stepwise hedonic regression. Besides the property characteristics,[12] we include a location variable and a dummy variable which is set equal one if the transaction is executed by an institutional investor, and the local GDP-level. As shown in Exhibit 3.3, all coefficients have the expected sign and all but the institutional-investor-dummy are statistically significant at the 1% level. Size and location of the condominium appear to have the strongest influence on the selling price.

Model II: The second model is supplemented by the out-of-town variable (lnDISTANCE). With this model, we aim to investigate whether out-of-town buyers pay higher prices for real estate, than their local counterparts.

The coefficient of the out-of-town variable is positive and statistically significant for both years. Hence, we show empirically that distant buyers indeed payed a premium for condominiums in Miami-Dade County in both years (Exhibit 3.3). The coefficient can be interpreted as follows: the further away the buyer lives, the higher the price he pays for a condominium. Our findings support the conclusions of Lambson *et al.* (2004), Ihlanfeldt and Mayock (2012), and Zhou *et al.* (2015), that the observed price premiums are explained by distant investors who face higher search costs and are informationally disadvantaged compared to local investors. The second important question we aim to answer with this model is whether the premium caused by distance decreases over time, due to a reduction of search costs and therefore better public information. We find that the out-of-town buyer coefficient (lnDISTANCE) decreases by 20.63% from 0.0572 in 2005 to 0.0454 in 2015. The coefficients translate into an out-of-town buyer paying approximately $5,000 more in 2005 than in 2015 if he lives 100 miles away. In conclusion, the results of *Model II* strongly support the hypothesis that reduced search costs due to better information availability lead to a smaller premium for out-of-town buyers in 2015 (Pereira, 2005).

While the out-of-town variable ln(DISTANCE) addresses the price differential paid by potentially less informed buyers, *Model II* does not account for price variations due to anchoring.

[12] Property characteristics are: square feet, number of bedrooms, number of bathrooms, and ln(age).

Exhibit 3.3 | Regression results Models I & II

	Model I			
	2005		2015	
	Coefficient	t-stat	Coefficient	t-stat
Intercept	12.0182***	408.75	12.1016***	338.69
SQFT	0.0009***	32.46	0.0011***	29.62
NBR BEDRMS	-0.1389***	-10.27	-0.2782***	-22.53
NBR BATH	0.0770***	4.37	0.1614***	9.72
ln(AGE)	-0.0958***	-17.48	-0.1450***	-20.21
LOCATION	-0.0160***	-23.84	-0.0236***	-42.67
INST INV	0.0442**	1.97	0.0735***	7.06
GDP	3.08E-07***	15.48	2.56E-07***	14.36
ln(DISTANCE)				
ANCHOR				
WEALTH				
Adjusted R^2	0.5417		0.5072	
Observations	4,431		9,378	

	Model II			
	2005		2015	
	Coefficient	t-stat	Coefficient	t-stat
Intercept	11.9000***	379.25	11.9864***	328.05
SQFT	0.0008***	30.30	0.0011***	41.43
NBR BEDRMS	-0.1200***	-7.87	-0.2816***	-23.07
NBR BATH	0.0662***	3.44	0.1641***	10.27
ln(AGE)	-0.0889***	-15.58	-0.1414***	-18.51
LOCATION	-0.0154***	-21.79	-0.0238***	-40.91
INST INV	0.0553***	2.30	0.0850***	7.66
GDP	9.52E-08***	3.92	1.25E-07***	6.07
ln(DISTANCE)	**0.0572*****	**14.82**	**0.0454*****	**13.39**
ANCHOR				
WEALTH				
Adjusted R^2	0.5649		0.5226	
Observations	3,053		6,918	

Notes: Exhibit 3.3 shows the regression results of *Models I* and *II*. The first model is the baseline, which is a conventional hedonic pricing regression. *Model II* includes the out-of-town variable (lnDISTANCE) which serves as a proxy for search costs. The dependent variable is the natural log of the sales price. All models use robust standard errors. *, **, and *** denote statistical significance at the 10, 5, and 1% levels, respectively.

Model III: In order to investigate whether buyers pay a premium due to biased beliefs, we enlarge our regression model by another variable (ANCHOR), which supposedly induces a premium as well. All of the coefficients are not only collectively, but also individually significant and have the expected sign, except for the institutional investor dummy, which is only significant in 2015, and GDP (Exhibit 3.4). The coefficients of ANCHOR are positive and statistically significant in both years at 0.0003. This result indicates that buyers from higher priced cities are paying significantly more than buyers from cities with lower housing prices – even if the coefficients are fairly small. The anchoring effect remains at the same level for both years. At the same time, the out-of-town indicator (lnDISTANCE) is still positive, statistically significant, and decreases from 2005 to 2015. These two coefficients together point to significant price premiums for properties sold to out-of-town buyers as a result of both information asymmetry and anchoring, with the anchoring effect being far smaller.

Exhibit 3.4 | Regression results Models III & IV

	Model III			
	2005		2015	
	Coefficient	t-stat	Coefficient	t-stat
Intercept	11.8654***	342.37	11.9306***	302.00
SQFT	0.0009***	26.86	0.0011***	38.70
NBR BEDRMS	-0.1142***	-6.88	-0.2753***	-20.91
NBR BATH	0.0381*	1.92	0.1713***	10.06
ln(AGE)	-0.0921***	-14.25	-0.1379***	-16.85
LOCATION	-0.0149***	-18.99	-0.0232***	-36.35
INST INV	0.0371	1.44	0.0953***	8.00
GDP	1.87E-08	0.50	3.24E-08	0.96
ln(DISTANCE)	**0.0466*** **	**12.11**	**0.0416*** **	**10.09**
ANCHOR	**0.0003*** **	**5.17**	**0.0003*** **	**7.95**
WEALTH				
Adjusted R^2	0.5561		0.5176	
Observations	3,053		6,918	

	Model IV			
	2005		2015	
	Coefficient	t-stat	Coefficient	t-stat
Intercept	11.8758***	337.86	11.9421***	294.16
SQFT	0.0009***	26.84	0.0011***	38.7
NBR BEDRMS	-0.1141***	-6.87	-0.2753***	-20.90
NBR BATH	0.0388*	1.95	0.1711***	10.06
ln(AGE)	-0.0920***	-14.25	-0.1384***	-16.87
LOCATION	-0.0149***	-19.03	-0.0231***	-36.27
INST INV	0.0376	1.46	0.0966***	8.09
GDP	4.41E-08	1.13	5.42E-08	1.42
ln(DISTANCE)	**0.0466*** **	**12.12**	**0.0401*** **	**9.49**
ANCHOR	**0.0003*** **	**4.77**	**0.0003*** **	**7.79**
WEALTH	**2.23E-06*** **	**2.11**	**2.68E-06**	**1.29**
Adjusted R^2	0.5567		0.5177	
Observations	3,053		6,918	

Notes: Exhibit 3.4 shows the results of regression *Models III* and *IV*. The third model includes both the out-of-town variable (lnDISTANCE) and the anchoring. The fourth model includes all three key variables of interest: lnDISTANCE, ANCHOR, and WEALTH. The dependent variable is the natural log of the sales price. All models use robust standard errors. *, **, and *** denote statistical significance at the 10, 5, and 1% levels, respectively.

Model IV: Another way to control for the heterogeneity of buyers, besides different information levels and different pricing beliefs, is to include a wealth variable, which refers to the average personal income per capita at the buyer's original location minus the average personal income per capita at the property's location. We hypothesize that buyers from areas with a higher personal per capita income are more likely to pay higher prices than others. This effect was mainly measured as part of the anchoring effect in existing literature.

Exhibit 3.4, which shows the results of *Model IV*, indicates that the wealth variable is statistically significant in 2005. Therefore, buyers from regions with a higher income level per capita pay higher prices for real estate even if the premium is fairly small. In 2015 this variable has no influence on the sales price, as the coefficient (WEALTH) is nearly 0 and not statistically significant. Hence, the personal income level can be regarded as another source of out-of-town buyer premiums but only for 2005. Other macroeconomic factors like CPI, employment, or net earnings have been tested and do not report any significant influence on the sales price either. Microeconomic variables, such as the buyer's individual personal income for each transaction, are supposed to yield more explanatory power but were not available. Nevertheless, the coefficients ln(DISTANCE) and ANCHOR are still positive and statistically significant, which supports the results of *Model III*.

3.5.2 Robustness Checks

Variations to Model IV

Exhibit 3.5 reports the results of ten additional robustness tests to validate the regression results by concentrating on the three variables of interest (lnDISTANCE, ANCHOR, and WEALTH). In each of the ten cases, we change one aspect of *Model IV* concerning the variables age (cases 1 & 2), selling price (cases 3 & 4), type of investor (cases 5 & 6), location (case 7), market value (case 8), or different out-of-town measures (cases 9 & 10).

In cases 1 & 2, we find that the out-of-town and anchoring premium occurs both for new (age < 5 years) and for older (age > 5 years) properties, whereas the out-of-town coefficient (lnDISTANCE) is not significant in 2015 for new properties. For both age categories, the out-of-town coefficient decreases from 2005 to 2015 and the anchoring effect appears to have a greater impact on the selling price, when compared to the results of the regular *Model IV*. The wealth effect has no influence on the selling price. With an adjusted R-squared of about 70% in both years, the clustering in age categories strongly supports our results of *Model IV*.

Exhibit 3.5 | Robustness checks

		Model IV									
		Distance		Anchoring		Wealth		Observations		R-Squared	
Case	Variation to (9)	2005	2015	2005	2015	2005	2015	2005	2015	2005	2015
1	Age < 5 years	0.0229***	0.0107	0.0053***	0.0018***	-3.46E-07	-3.40E-06	761	215	0.7178	0.6958
2	Age > 5 years	0.0120***	0.0049*	0.0039***	0.0037***	1.82E-06	-1.75E-07	2,989	7,543	0.7316	0.7122
3	High Price	0.0096***	0.0131***	0.0031***	0.0014***	2.33E-06	-3.20E-07	1,146	2,178	0.6371	0.3918
4	Low Price	0.0016	-0.0034	0.0024***	0.0027***	8.18E-08	-1.77E-06	2,604	5,580	0.3886	0.5413
5	Institutional	0.0617***	0.0224***	0.0004***	0.0004***	-1.71E-06	6.93E-06	374	4,267	0.5650	0.4815
6	Non-Insitutional	0.0453***	0.0606***	0.0002***	0.0002***	5.18E-06**	2.58E-07	2,679	2,651	0.5543	0.5942
7	Miami City	0.0178***	-0.0079	0.0002**	0.0002***	7.97E-07	6.58E-07	1,387	3,320	0.6056	0.6263
8	Market Value included	0.0094***	0.0005	-0.0001	0.0001***	4.18E-07	3.35E-07	3,523	7,574	0.7977	0.8460
9	Out-of-state dummy	0.2742***	0.2659***	0.0004***	0.0005***	1.71E-06	-1.86E-07	3,635	8,267	0.5568	0.5218
10	Out-of-town dummy	0.1093***	0.0116	0.0004***	0.0005***	4.32E-06*	6.32E-06***	3,635	8,267	0.5481	0.5159

Notes: This table shows the coefficient results of the three variables of interest (lnDISTANCE, ANCHOR and WEALTH) for 10 different modifications to Model IV for the years 2005 and 2015. The results of all variables are available upon request. Case 1: includes only properties with age < 5 years; case 2: includes only properties with age > 5 years; case 3: includes only properties with selling price higher than the mean; case 4: includes only properties with selling price lower than the mean; case 5: includes only transactions made by institutional investors; case 6: includes only transactions made by individual investors; case 7: includes only properties in Miami City; case 8: includes estimated market values instead of property characteristics; case 9: includes out-of-state dummy instead of distance, which is set to 1 if the buyer is from a state other than Florida; case 10: includes out-of-town dummy instead of distance, which is set to 1 if the buyer lives more than 50 miles away from the property he buys. Only the variables of interest (distance, anchoring and wealth) are shown in this table. Dependent variable is the natural log of the sales price. All models use robust standard errors. * Indicates significance at the 10% level. ** Indicates significance at the 5% level. *** Indicates significance at the 1% level.

Furthermore, we divide the sample into high price (with a selling price above average) vs. low price (with a selling price below average) properties and into transactions by institutional vs. non-institutional investors. In all cases (3-6), the out-of-town and anchoring coefficients are statistically significant except for properties with lower prices.[13] The out-of-town coefficient even shows a negative sign in 2015 for properties with selling prices below the average, which indicates that with increasing distance, the buyer pays less for the property. The same phenomenon of a negative out-of-town coefficient for 2015 can be observed when considering only Miami City.

For the different types of investors (case 5 & 6) both distance and anchoring are statistically significant at the 1% level. The wealth coefficient is statistically significant in 2005 but not in 2015.

Analyzing potential distance, behavioral, and wealth effects, it is inferentially critical to control carefully for property and transaction attributes. The estimated price differential associated with distant buyers could be biased, because of omitted quality characteristics correlated with distance. In order to control for potentially omitted variables, we use an independent estimate of the market value in our regression, following Ihlanfeldt and Mayock (2012). They state that including the estimated market value in a hedonic regression obviates the need to include any property descriptors, since this value summarizes the locational and structural characteristics into a single number. By including the market value instead of housing characteristics, the models reach an adjusted R-squared of nearly 80%. The out-of-town coefficient is positive and statistically significant in 2005. It decreases to 0.0005 in 2015, but is not statistically significant. Therefore, this result is still in line with our hypothesis that the out-of-town buyer premium is supposed to decline. The anchoring coefficient is negative and not statistically significant in 2005 but positive and significant in 2015. Hence, it seems that anchoring plays a greater role in 2015 compared to 2005, whereas for distance, it is the other way round.

In case 9, we repeat the regression analysis of *Model IV* and use an out-of-state dummy variable[14] instead of a continuous out-of-town variable (lnDISTANCE). The results strongly support our findings from *Model IV* for all three variables of interest.

Another variation to the out-of-town variable is tested in case 10. Here, we use an out-of-town dummy[15] instead of distance in miles. In 2015, the out-of-town dummy coefficient is not statistically significant but smaller than in 2005. With this modification, the wealth coefficient is positive and statistically significant for both years for the first time but still very small.

[13] With one exception: case 4, ANCHOR in 2005.
[14] The out-of-state dummy is set to one if the buyer does not live in Florida and zero otherwise.
[15] The out-of-town dummy is set to one if the buyer lives more than 50 miles away from the property and zero otherwise.

None of these additional robustness tests changes the answer to our questions of whether out-of-town buyers pay a premium, what causes the premium, and whether the premium caused by distance declines from 2005 to 2015.

Propensity Score matching

In Exhibit 3.1 it can be seen that the properties sold in 2005 are fundamentally different from the properties sold in 2015 in terms of their characteristics. In order to ensure that the difference in out-of-town buyer premiums over the years does not result from comparing different properties in 2005 and 2015, a propensity score matching is conducted. This statistical technique attempts to reduce the selection bias due to confounding variables that could be found in an estimate of the treatment group versus the group that did not receive the treatment. In identifying appropriate properties for the matched sample, we deploy the propensity score matching algorithm which was introduced by Rosenbaum and Rubin (1983).

The basic idea is to find those transactions in a large control group which are similar to the transactions in the treatment group in terms of the relevant housing characteristics. In our case, 2005 is the treatment group and 2015 is the control group as the number of transactions is larger in 2015 than in 2005. The propensity score matching offers the advantage of matching various characteristics simultaneously and thus reducing the multidimensionality of the matching parameters into one single measure, namely the propensity score (Li and Zhao, 2006).

Based on the likelihood estimation, the propensity score is calculated from a logit discrete-choice model. Having estimated the propensity score for all properties in 2005 and 2015, we follow Hillion and Vermaelen (2004) and apply the nearest neighbor matching algorithm, which determines the matching partner in the control group (2015) on the basis of the closest propensity score. This approach enables us to identify the statistical "2015 twin" to each transaction in 2005 in terms of relevant housing characteristics.

Based on equivalent counterparts of transactions conducted in 2015, we create a matched sample, while ensuring an appropriate matching quality. Finally, the matched sample is used to run regression *Model IV* for the years 2005 and 2015 in order to check if the out-of-town buyer premium is still present over time.

The results of the matched sample, as shown in Appendix 3.5, indicate that out-of-town buyers pay a statistically significant premium due to higher search costs in 2005 and 2015. This premium decreases from 0.0466 to 0.0149 from 2005 to 2015. The anchoring coefficient is statistically significant for both years at 0.0003. The average level of the personal income per capita plays a significant role only in 2005. Overall, the results of the matched sample strongly support our regression results as the pattern of coefficient signs and statistical significance proves to be true.

Intercounty Comparison

As another robustness check, we run the regression on *Model IV* for two other counties – San Diego and San Francisco.[16] Previous research has either focused on one city or region or even pooled the transactions of several metropolitan areas in order to reach a reasonable sample size per property type (see Appendix 3.1). Our study extends the existing literature on out-of-town models by considering more than one county. With this intercounty comparison, we investigate whether the regression results and conclusions for Miami-Dade County are applicable to other counties in the US as well. The focus of this intercounty comparison is on the estimated coefficients of ln(DISTANCE), ANCHOR, and WEALTH, as they are our key variables of interest and represent the pricing effects associated with search costs, anchoring, and the personal income level.

Exhibit 3.6 contains the results from estimating *Model IV* for Miami-Dade County, San Diego County, and San Francisco County. Before running the regressions for San Diego and San Francisco County, we implement the same four data-cleaning steps as for Miami-Dade.[17]

In our final sample for San Diego, we identify 961 properties in 2005 and 1,224 in 2015. The adjusted R-squared for the models are 52% and 50%. The out-of-town variable ln(DISTANCE) is positive and highly significant at 0.0291 in 2005 and 0.0693 in 2015. The results strongly suggest that out-of-town buyers pay a premium, but this premium increased from 2005 to 2015. Therefore, our hypothesis that better information availability leads to lower premiums cannot be confirmed for San Diego County. The anchoring coefficient yields almost the same results for San Diego County compared to Miami-Dade County. Hence, the anchoring effect is validated with these results – the buyer's beliefs appear to exert an impact on the selling prices.

Surprisingly, in San Diego County, the WEALTH-coefficient is statistically significant at the 1% level and yields a very small but negative number. In San Diego County, the personal income level per capita of the buyer's origin obviously influences the selling price negatively.

[16] We choose these two counties due to data availability and in order to compare the eastern and western parts of the US.

[17] 1) Non-owner-occupied condominiums only. 2) Transaction records missing one or more variables that are required for regression, are excluded. 3) Only using p5-p95 quantile for price per square foot and selling price. 4) Properties with more than four bedrooms and more than six bathrooms are excluded, as well as the combination of zero bed and bathrooms.

Exhibit 3.6 | Intercounty comparison

	MIAMI DADE COUNTY - Model IV				*SAN DIEGO COUNTY* - Model IV				*SAN FRANCISCO COUNTY* - Model IV			
	2005		2015		2005		2015		2005		2015	
	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Intercept	11.8758***	337.86	11.9421***	294.16	12.0576***	157.68	11.7318***	72.32	13.0680***	139.76	13.3338***	176.42
SQFT	0.0009***	26.84	0.0011***	38.70	0.0007***	14.60	0.0007***	5.73	0.0003***	5.92	0.0003***	12.65
NBR BEDRMS	-0.1141***	-6.87	-0.2753***	-20.90	-0.0674***	-3.70	-0.1167***	-3.66	0.0818***	4.97	0.1311***	7.07
NBR BATH	0.0388*	1.95	0.1711***	10.06	0.0828***	3.15	0.1866***	3.74	-0.0095	-0.26	0.0244	1.18
ln(AGE)	-0.0920***	-14.25	-0.1384***	-16.87	-0.0199	-1.54	-0.2066	-0.53	0.0074	0.67	-0.0310	-0.31
LOCATION	-0.0149***	-19.03	-0.0231***	-36.27	-0.0043***	-6.13	-0.0031***	-3.38	-0.0165**	-2.08	-0.0286***	-4.29
INST INV	0.0376	1.46	0.0966***	8.09	0.0861***	3.52	-0.0828***	2.86	0.0393	1.10	0.0204	0.67
GDP	4.41E-08	1.13	5.42E-08	1.42	1.34E-07***	2.01	-1.11E-07**	-2.09	1.05E-07	1.07	6.48E-09	-0.09
ln(DISTANCE)	**0.0466*****	**12.12**	**0.0401*****	**9.49**	**0.0291*****	**5.60**	**0.0693*****	**10.08**	**0.0087**	**0.67**	**0.0301*****	**3.04**
ANCHOR	**0.0003*****	**4.77**	**0.0003*****	**7.79**	**0.0002****	**1.65**	**0.0004*****	**4.46**	**-0.0001**	**-1.22**	**8.69E-06**	**0.13**
WEALTH	**2.23E-06*****	**2.11**	**2.68E-06**	**1.29**	**-8.89E-06*****	**-3.74**	**-8.00E-06*****	**-5.19**	**6.99E-08**	**0.02**	**8.95E-07**	**0.55**
Adjusted R^2	0.5567		0.5177		0.5239		0.4979		0.6017		0.5739	
Observations	3,053		6,918		961		1,224		220		500	

Notes: The exhibit shows the regression results for Model IV for Miami-Dade County, San Diego County and San Francisco County in the years 2005 and 2015. Dependent variable is the natural log of the sales price. All models use robust standard errors. * Indicates significance at the 10% level. ** Indicates significance at the 5% level. *** Indicates significance at the 1% level.

The sample for San Francisco County contains 220 properties in 2005 and 500 in 2015. The sample size is rather small compared to Miami-Dade or San Diego County, but all three counties have increasing sample sizes in 2015, due to better data availability. The San Francisco Model explains 60% of the variation in logged sales prices in 2005 and 57% in 2015. The out-of-town variable reports positive coefficients in both years, but only 2015 is statistically significant and higher than 2005. Therefore, we have similar results as in San Diego County – distant buyers pay more for real estate in 2015 than in 2005. The anchoring coefficient is not statistically significant but negative in 2005 and suggests that biased beliefs of buyers due to different price levels do not cause serious price differentials. Furthermore, the wealth effect does not influence the selling price, as the coefficient is nearly zero and not statistically significant.

Some reasons why we might obtain different results in three different markets include the possibilities that (1) market participants are more or less familiar with some markets, for example, where they spend more time on holiday or in vacation rentals, (2) the inventory levels vary by market and price range and higher priced markets generally require more time on the market, (3) more or less elastic barriers to adding new supply may influence anticipated price increases, (4) the presence of a higher proportion of foreign or institutional buyers may influence our results, (5) especially San Francisco is somehow unique as it represents a market where global firms acquire highly skilled human capital almost disregarding price levels. These issues are worthy of further study but require detailed information on buyers, and such data were not available here.

With the intercounty comparison, we aim to show that results for one regional market cannot simply be transferred to other housing markets in the US. The hypothesis of smaller out-of-town buyer premiums due to better public information has been shown to hold for Miami-Dade County but could not be confirmed for San Diego or San Francisco County. The reasons for varying results for different counties can be attributed to heterogeneous regional markets and buyer characteristics like taxation, market elasticity, labor demand, and experience. By employing the same methodology for two other counties, this paper makes a contribution by explicitly pointing out an important caveat that has often been ignored in the previous literature. That is, real estate prices do not solely reflect housing characteristics but are strongly influenced by macroeconomic and behavioral factors that are in some ways hard to measure.

3.6 Conclusion

The inefficiencies of the real estate market lead to disadvantaged market participant groups that pay higher prices for equivalent properties (Zhou *et al.*, 2015). This disadvantage often arises from asymmetric information and different search costs among potential buyers.

Out-of-town buyers are considered to represent an informationally disadvantaged group, as conventional wisdom posits that they incur higher search costs and therefore pay a premium for real estate, compared to local buyers. Although the pricing of investments by distant vs. local buyers is likely to be affected by search costs, behavioral bias, and income level, the relative importance of these effects has not been examined comprehensively in the literature.

Empirically, we address three questions concerning the out-of-town buyer premium. First, do out-of-town buyers overpay for real estate? Using a sample of 3,053 condominium transactions in 2005 and 6,918 in 2015 in Miami-Dade County, we find that out-of-town buyers do pay a positive and statistically significant premium, compared to local buyers in both years. In 2005, distant buyers payed approximately $5,000 more for a condominium than in 2015 if they live 100 miles away.

Second, is the out-of-town buyer premium caused by higher search costs/information asymmetries (distance), biased beliefs (anchoring), or different income levels (wealth)? We use distance as a continuous out-of-town variable, because it is supposed to be a good proxy for search costs, and we find that distance causes a statistically significant premium. Biased beliefs in the form of anchoring seem to play a less important role, as the premium is very small but still statistically significant. The personal income level has only statistical significant impact on the selling price in 2005. In conclusion, we find that the largest premium paid by out-of-town buyers is due to information asymmetries (distance). Supposedly, out-of-town buyers face higher search costs and therefore may decrease the number of searches. The consequence is that they evidently pay more than better informed local buyers.

Third, did the out-of-town buyer premium caused by information asymmetries (distance) decline over time? Two points in time (2005 and 2015) with huge differences in information availability are considered in order to investigate whether the internet now successfully ameliorates the differences in information level between out-of-town and in-town buyers over time. We use distance as a proxy for search costs and therefore information levels for buyers and find evidence that the out-of-town buyer premium due to distance decreased by 20.63% from 2005 to 2015. Therefore, the internet likely contributes to equalizing prices paid by out-of-town buyers and locals to some extent, as the premium in 2015 is not as high as in 2005. But real estate remains a far less than perfectly competitive market characterized by price dispersion, high transactions costs, and noise. There may be further legitimate doubts as to whether the decreasing premium caused by distance from 2005 to 2015 is solely related to increased information availability through the internet. We cannot fully resolve this concern as other factors of the housing market may have changed between 2005 and 2015, which we were not able to control for due to data availability issues. This includes buyer locality and quality composition, changes in housing stock, exchange rates, and regulations among other things. Hence, we cannot conclude yet that greater information availability through the internet fully equalizes the information level between heterogeneous buyers as out-of-town buyers still have higher search costs in 2015 than local buyers.

The propensity score matching and various robustness checks validate our results. With the intercounty comparison, we show that further work is needed in order to better isolate the causes of the out-of-town buyer premium in real estate markets.

Our research is limited to condominium transactions in Miami-Dade County in 2005 and 2015. Alternative types of real estate and other geographic areas need to be examined in detail. Data for other regions and additional variables such as detailed information about additional housing amenities, homeowner association fees, macroeconomic factors, time on market, or data for owner-occupied properties, would be a good extension to the model but were not available. Furthermore, the role of real estate agents as information intermediaries could usefully be included in further research, as they evidently still play an important role for the housing market.

3.7 Appendix

Appendix 3.1 | Relevant literature

Paper	Location	Time	Out-of-town buyer type	Dependent variable	Out-of-town variable (Coefficient)	Anchoring variable (Coefficient)	Observations	Methodology
Miller; Rice (1981)	Columbus, Ohio	1976	Own usage	Selling Price	dummy (3.179)	-	91	OLS
Miller; Sklarz; Ordway (1988)	Honolulu, Hawaii	1986-1988	n.a.	Selling Price	dummy (11.11)	-	421	OLS
Turnbull; Sirmans (1993)	Baton Rouge, Louisiana	1988-1989	Own usage	ln(Selling Price)	dummy (0.054)	-	151	OLS
Watkins; Craig (1998)	Glasgow, Scottland	1991-1992	Own usage	Selling Price	separate models	-	544	OLS
Lambson; McQueen; Slade (2004)	Phoenix, Arizona	1990-2002	Investment	ln(price per unit)	dummy (0.0537)	median price (0.0506)	2,854	OLS
Clauretie; Thistle (2007)	Las Vegas, Nevada	2000-2004	Investment	Price & Time on market	distance (0.0039)	price difference (0.5131)	2,828	TSLS
Ihlanfeldt; Mayock (2012)	Florida	2009	Own usage	ln(Selling Price)	ln(distance) (0.003)	ln(ppsqft) (0.049)	6,779	OLS
Zhou; Gibler; Zahirovic-Herbert (2015)	Chengdu, China	2013	Investment/ Own usage	ln(price)	ln(distance) (0.048)	average price (0.052)	906	OLS
Ling, Naranjo, Petrova (2016)	15 Metropolitan Areas	1997-2011	Investment	ln(price)	dummy (0.1219)	dummy (0.0034)	48,318	OLS
Horenstein, Osgood, Snir (2017)	Yavapi County, Arizona	1991-2000	Investment	annualized returns	dummy (0.10)	-	1,013	OLS

Notes: The table gives an overview of the most relevant papers related to the out-of-town buyer premium topic. The papers are ranked chronologically.

Appendix 3.2 | Variable description

$P \in \{0,\infty\}$	Sales price
$P* \in \{0,\infty\}$	Optimal reservation price
$SC \in \{0,\infty\}$	Search costs
$V \in \{0,\infty\}$	Exogenous value of the property
$A \in \{-\infty;\infty\}$	Shift of probability weight on a priori distributions (Anchoring)

With ∞ stands for any sort of reservation level or overall income.

Equation (5) consists of the following parts:

$(V - P^*)$	=	Expected surplus of one additional search.
$[1 - F(P^*, A)](V - P^*)$	=	With the probability of *(1-F(P*,A)),* the buyer draws a sales price that is too high (*P* > *P**) and continues searching.
$\int_0^{P^*} (V - P)F(P, A)dP - SC$	=	Buyer draws a price lower than $P*$, stops searching, pays the price *P* and the search costs, and earns the surplus.

Appendix 3.3 | Correlation matrix

	Sales price	Sqft	Nbr bdrms	Nbr bath	Age	Location	Inst. Inv.	GDP	Distance	Anchor	Wealth
Sales price	1.00										
Sqft	0.60	1.00									
Nbr bdrms	0.24	0.72	1.00								
Nbr bath	0.46	0.75	0.74	1.00							
Age	-0.27	-0.31	-0.32	-0.29	1.00						
Location	-0.22	0.20	0.36	0.23	-0.06	1.00					
Inst. Inv.	0.10	0.04	0.03	0.06	0.02	-0.06	1.00				
GDP	0.14	0.05	-0.06	0.01	0.06	-0.02	-0.11	1.00			
Distance	0.19	0.08	-0.06	0.02	0.02	-0.05	-0.21	0.52	1.00		
Anchor	0.22	0.02	-0.12	-0.02	0.03	-0.21	0.01	0.61	0.34	1.00	
Wealth	0.12	0.04	-0.04	0.01	0.04	-0.03	-0.10	0.63	0.46	0.46	1.00

Appendix 3.4 | Comparison of information levels

Map which includes:	Information in 2005	Information in 2015
all homes for sale in a certain area	x	✓
comparable home prices	os	✓
auctions	os	✓
crime	os	✓
schools	os	✓
commuting	os	✓
shop & eat	os	✓
affordability	os	✓
age statistics	os	✓
hazards	os	✓
traffic	os	✓
street view	x	✓
photos	x	✓
Information about housing characteristics:		
sales price	✓	✓
address	✓	✓
number of bedrooms	✓	✓
number of bathrooms	✓	✓
square feet	✓	✓
estimated mortgage	✓	✓
visiting time	✓	✓
housing type (single family, condominium,…)	✓	✓
year built	✓	✓
last remodel year	✓	✓
heating/cooling	✓	✓
parking	✓	✓
price per sqft	✓	✓
view	x	✓
pool	✓	✓
hot tub/spa	✓	✓
barbecue area	✓	✓
stories	✓	✓
yard	✓	✓
lot size	✓	✓
features (stall shower, tub, granite countertops, fireplace)	✓	✓

	amenities (tennis area, biking, jogging, hiking area)	✓	✓
	utilties (sprinkler system)	✓	✓
	virtual tour	x	✓
Price and neighborhood information:			
	Zestimate with Zestimate history	x	✓
	price of comparable homes	os	✓
	local property tax assessments	os	✓
	local listing and sales prices	os	✓
	price history	x	✓
	price trends	os	✓
	tax history	os	✓
	mortgage calculator (affordability)	x	✓
	nearby similar sales	os	✓
	similar homes for sale	x	✓
	neighborhood price calculator with median Zestimate and "market temperature"	x	✓
	neighborhood map with prices and pictures of other homes	x	✓
	nearby schools with school rating	os	✓
	agent contact details	os	✓

✓ = available
x = not available
os = available through other sources

Appendix 3.5 | Regression results of Model IV (Matched Sample)

	Model IV			
	2005		2015	
Matched Sample	Coefficient	t-stat	Coefficient	t-stat
Intercept	11.8758	337.86	11.9148***	152.95
SQFT	0.0009***	26.84	0.0011***	21.40
NBR BEDRMS	-0.1141***	-6.87	-0.1973***	-7.42
NBR BATH	0.0387*	1.95	0.0969***	2.65
ln(AGE)	-0.0920***	-14.25	-0.1414***	-8.51
LOCATION	-0.0149***	-19.03	-0.0233***	-21.73
INST INV	0.0376	1.46	0.034546	1.14
GDP	-4.41E-08	-1.13	-9.94E-08	-1.48
ln(DISTANCE)	**0.0466*** **	**12.12**	**0.0149*** **	**6.03**
ANCHOR	**0.0003*** **	**4.77**	**0.0003*** **	**4.98**
WEALTH	**4.71E-6** **	**2.11**	**3.13E-06**	**1.13**
Adjusted R^2	0.5567		0.6018	
Observations	3,053		1,351	

3.8 References

Agarwal, S., Ben-David, I. and Yao, V. (2015), “Collateral Valuation and Borrower Financial Constraints. Evidence from the Residential Real Estate Market”, *Management Science*, Vol. 61 No. 9, pp. 2220–2240.

Biswas, D. (2004), “Economics of information in the Web economy”, *Journal of Business Research*, Vol. 57 No. 7, pp. 724–733.

Byrnjolfsson, E. and Smith, M.D. (2000), “Frictionless Commerce? A Comparison of Internet and Conventional Retailers”, *Management Science*, Vol. 46 No. 4, pp. 563–585.

Clauretie, T.M. and Thistle, P.D. (2007), “The Effect of Time-on-Market and Location on Search Costs and Anchoring. The Case of Single-Family Properties”, *The Journal of Real Estate Finance and Economics*, Vol. 35 No. 2, pp. 181–196.

Diaz, J. and Hansz, J.A. (1997), “How valuers use the value opinions of others”, *The Journal of Property Valuation and Investment*, Vol. 15 No. 3, pp. 256–260.

Giraud, T., Cura, R. and Viry, M. (2017), “Interface between R and the OpenStreetMap-Based Routing Service OSRM”.

Griffin, J.M. and Maturana, G. (2016), “Who Facilitated Misreporting in Securitized Loans?”, *Review of Financial Studies*, Vol. 29 No. 2, pp. 384–419.

Hillion, P. and Vermaelen, T. (2004), “Death spiral convertibles”, *Journal of Financial Economics*, Vol. 71 No. 2, pp. 381–415.

Hollas, D.R., Rutherford, R.C. and Thomson, T.A. (2010), “Zillow's Estimate of Single-Family Housing Values”, *The Appraisal Journal*, Winter, pp. 26–32.

Horenstein, A.R., Osgood, D. and Snir, A. (2017), “Out-of-Town Buyers, Mispricing, and the Availability Heuristic in a Housing Market”, *Real Estate Finance*, Summer, pp. 27–35.

Ihlanfeldt, K. and Mayock, T. (2012), “Information, search, and house prices: Revisited”, *The Journal of Real Estate Finance and Economics*, Vol. 44 No. 1-2, pp. 90–115.

Kohle, D. and Wickham, H. (2013), “ggmap: Spatial Visualization with ggplot2”, *The R Journal*, Vol. 5 No. 1, pp. 144–161.

Lambson, V.E., McQueen, G.R. and Slade, B.A. (2004), “Do Out-of-State Buyers Pay More for Real Estate? An Examination of Anchoring-Induced Bias and Search Costs”, *Real Estate Economics*, Vol. 32 No. 1, pp. 85–126.

Li, X. and Zhao, X. (2006), “Propensity score matching and abnormal performance after seasoned equity offerings”, *Journal of Empirical Finance*, Vol. 13 No. 3, pp. 351–370.

Ling, D.C., Naranjo, A. and Petrova, M.T. (2016), “Search Costs, Behavioral Biases, and Information Intermediary Effects”, *The Journal of Real Estate Finance and Economics*, pp. 1–38.

Lisi, G. (2014), “Home-seekers in the Housing Market”, *International Real Estate Review*, Vol. 17 No. 1, pp. 47–62.

Maury, T.-P. and Tripier, F. (2014), “Search strategies on the housing market and their implications on price dispersion”, *Journal of Housing Economics*, Vol. 26, pp. 55–80.

Miller, N.G. and Rice, P. (1981), “A Note on the Differentiation of the Market Participants by Search Costs”, *Economics Letters*, Vol. 7 No. 1, pp. 39–46.

Miller, N.G., Sklarz, M.A. and Ordway, N. (1988), “Japanese Purchases, Exchange Rates and Speculation in Residential Real Estate Markets”, *The Journal of Real Estate Research*, Vol. 3 No. 3, pp. 39–49.

Northcraft, G.B. and Neale, M. (1987), “Expert, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions”, *Organizational Behavioral and Human Decision Processes*, Vol. 39 No. 1, pp. 84–97.

Pereira, P. (2005), “Do Lower Search Costs Reduce Prices and Price Dispersion”, *Information Economics and Policy*, Vol. 17 No. 1, pp. 61–72.

Rosenbaum, P.R. and Rubin, D.B. (1983), “The central role of the Propensity Score in observational studies for causal effects”, *Biometrika*, Vol. 70 No. 1, pp. 41–55.

Rothschild, M. (1974), “Searching for the Lowest Price When the Distribution of Prices Is Unknown: A Summary”, *Annals of Economic and Social Measurement*, Vol. 3 No. 1, pp. 293–294.

Smith, G.E., Venkatraman, M.P. and Dholakia, R.R. (1999), “Diagnosing the search cost effect: Waiting time and the moderating impact of prior category knowledge”, *Journal of Economic Psychology*, Vol. 20 No. 3, pp. 285–314.

Stigler, G.J. (1961), “The Economics of Information”, *The Journal of Political Economy*, Vol. 69 No. 3, pp. 213–225.

Turnbull, G.K. and Sirmans, C.F. (1993), “Information, search, and house prices”, *Regional Science and Urban Economics*, Vol. 23 No. 4, pp. 545–557.

Tversky, A. and Kahneman, D. (1974), “Judgment under Uncertainty: Heuristics and Biases”, *Science*, Vol. 185 No. 4157, pp. 1124–1131.

Watkins, C. (1998), “Are new entrants to the residential property market informationally disadvantaged?”, *Journal of Property Research*, Vol. 15 No. 1, pp. 57–70.

Weirick, W. and Ingram, F. (1990), “Finctional From choice in Applied Real Estate Analysis”, *Appraisal Journal*, January, pp. 57–73.

Zhang, Y., Zhang, H. and Seiler, M.J. (2016), “The Impact of Information Disclosure on Price Fluctuations and Housing Bubbles: An Experimental Study”, *Journal of Housing Research*, Vol. 25 No. 2, pp. 171–193.

Zhou, X., Gibler, K. and Zahirovic-Herbert, V. (2015), “Asymmetric buyer information influence on price in a homogeneous housing market”, *Urban Studies*, Vol. 52 No. 5, pp. 891–905.

Zilllow (2017), https://www.zillow.com/corp/About.htm.

4 Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany

Abstract

Purpose – The purpose of this study is to examine the value of text-based sentiment analysis for German real estate markets. By developing the first *German Real Estate Sentiment Dictionary,* this paper lays the foundation for future real estate-related textual analysis applications in Germany.

Design/Methodology/Approach – Conducting a large survey among about 1,700 real estate professionals, enables generating the *German Real Estate Sentiment Dictionary* with objective sentiment scores for real estate-related German words. Accordingly, this paper examines 125,462 newspaper articles published in the *Immobilien Zeitung,* the major real estate news provider in Germany, by applying the dictionary-based approach. A vector autoregressive framework and out-of-sample forecasts are utilized to analyze the dynamic relationship between news-based sentiment measures and the German residential market from 2007 to 2017.

Findings – Overall, the results yield strong and robust evidence of a significant relationship between the extracted sentiment and the housing market. More precisely, the negative sentiment indicator Granger-causes one-month-ahead *IMX* returns, even when controlling for macroeconomic fundamentals and an indirect sentiment measure. Furthermore, this paper finds that the analysis of German newspaper headlines alone, and analyzing the complete article, both constitute significant real estate sentiment measures.

Originality/Value – Most notably, in this paper, an objectively validated *German Real Estate Sentiment Dictionary* with 14,137 real estate-related words is developed. This exceptional resource will enable future researchers, as well as industry participants, to analyze all kinds of German text documents with regard to their inherent real estate sentiment. For the first time, sentiment measures from German real estate-related news items were extracted and subsequently applied to the market.

4.1 Introduction

Germany is Europe's strongest and largest economy and ranks fourth in the world after the United States, China, and Japan in terms of nominal GDP (International Monetary Fund, 2017). Within the German economy, real estate is one of the largest industries at about 18% gross value added. Furthermore, real estate comprises 80% of gross invested assets and is hence the most important asset class (Just *et al.*, 2017). Although Germany is one of the world's leading real estate markets, the market still lags behind global players such as the UK and the US in terms of data availability, market transparency and academic research (Maurer *et al.*, 2004; Schulte *et al.*, 2005). According to Schulte *et al.* (2005), academic and general research institutions play an important role in improving real estate market transparency in Germany. Comparing global real estate market transparency levels, JLL (2016) reports that Germany has improved remarkably over the last decade. Nevertheless, in the field of national real estate research, Germany still has a need to catch up, in order to gain deeper insights into real estate market dynamics. Especially contemporary approaches such as sentiment analysis have not yet found their way into German real estate research.

Until now, international research has consistently confirmed the value of sentiment for explaining real estate market movements (Clayton *et al.*, 2009; Ling *et al.*, 2014; Marcato and Nanda, 2016; Freybote and Seagraves, 2017). Within this field, textual analysis has recently attracted much attention from academia. The digitization of all kinds of text corpora and technical advances have opened up a potential new field of sentiment analysis, namely textual analysis. Various information- and sentiment-bearing texts such as news, earnings press releases, annual reports, 10 Ks, analyst reports, commentaries, or IPO prospectuses are now available online and can be analyzed with innovative textual analysis techniques.

The first economically relevant textual-analysis studies were conducted in the field of finance. They demonstrate that the "tone" extracted from various text documents contains information relevant to future stock market returns (Tetlock, 2007; Tetlock *et al.*, 2008; Bollen *et al.*, 2011) or trading volumes (Price *et al.*, 2012). Some initial attempts to apply this methodology to the real estate sector were conducted by Walker (2014), who found that media Granger-caused real house-price changes between 1993 and 2008 in the UK. Thereafter, Soo (2015) analyzed 34 cities across the US and confirmed earlier findings by showing that measures of housing sentiment predict future house prices. Investigating about 125,000 real-estate-related newspaper article headlines, Ruscheinsky *et al.* (2018) provide evidence of a leading relationship between media tone and future US REIT market movements.

However, these innovative textual analysis approaches have not yet been applied to German real estate markets. Reasons for this might include difficulties in accessing data and language barriers.

Among other things, the dictionary-based approach cannot be applied without a sentiment dictionary in the local language. Until now, there have been no efforts – to the authors' best knowledge – to establish a German sentiment dictionary for economic contexts.

Hence, the purpose of this paper is to develop the first *German Real Estate Sentiment Dictionary*, which makes it possible to apply the dictionary-based approach to German real estate-related news. Following Loughran and McDonald (2011, 2015), who found that discipline-specific pre-annotated word lists lead to better classification results than general ones, the aim is to include words regarding their meaning in a real estate context only. Conducting a large survey among about 1,700 real estate professionals, enables generating a *German Real Estate Sentiment Dictionary* with objective sentiment scores for 14,137 real estate-related German words. Having this exceptional resource available, allows analyzing all kinds of text documents with regard to their inherent real estate sentiment.

Accordingly, this paper examines 125,462 newspaper articles from 2007 to 2017, published in the *Immobilien Zeitung,* the major real estate news provider in Germany (Edelmann.ergo, 2017), by applying the dictionary-based approach. Aggregating the thus gained sentiment classifications enables calculating monthly positive and negative sentiment measures. As residential property represents the largest share in the property industry, with approximately 60% of total net fixed assets (Just and Maennig, 2012), the *IMX* apartment price index from *ImmobilienScout24* is selected. Due to the fact that direct real estate markets tend to be less transparent, one could expect sentiment to be even more relevant. Accordingly, the dynamic relationship between news-based sentiment measures and the residential real estate market is investigated in a vector autoregressive framework. Out-of-sample forecasts of direct real estate market returns complete the analysis.

The results yield strong and robust evidence of a significant relationship between news-based sentiment measures and housing market movements. More precisely, the negative sentiment indicator has predictive power over future *IMX* returns, even when controlling for macroeconomic fundamentals. This is even the case when controlling for another sentiment measure in the vector autoregressive model. Sentiment measures generated using the *German Real Estate Sentiment Dictionary* reveal superior results, compared to those constructed with general dictionaries such as *SentiWS* or *German Polarity Clues*. Furthermore, this paper found that the analysis of newspaper headlines alone, and analyzing the complete article, both constitute significant real estate sentiment measures. Constructing and utilizing different scopes for design of the *German Real Estate Sentiment Dictionary* confirmed earlier research findings regarding the supreme importance of using an appropriate sentiment dictionary. The comparison of forecasting accuracy further supports the added value of taking sentiment measures into consideration.

The findings provide insights that enable a better understanding of influences on German residential market movements that are not based solely on fundamental value changes. Most notably, an objectively validated *German Real Estate Sentiment Dictionary* is developed, which lays the foundation for future textual analysis in the German real estate market. Furthermore, this paper is the first to compare various dictionary designs in order to identify which yields the highest predictive power. Novel insights are ascertained by investigating different parts of newspaper articles. Accordingly, this paper makes a valuable contribution to the emerging literature on textual analysis and takes the first step for future applications in the German real estate market.

The remainder of the paper is structured as follows. The second section reviews the relevant literature in the field of sentiment extraction from textual data and its application to real estate markets. The third section describes the creation of the *German Real Estate Sentiment Dictionary*, which comprises three sequential steps. Data description and summary statistics are presented in the fourth section, while the fifth contains the methodology for the dictionary-based approach, the sentiment measure creation, and the vector autoregressive framework. The results are presented and discussed in the sixth section, together with an evaluation of the forecasting accuracy. The seventh section provides numerous robustness tests. Conclusions and practical implications of the developed *German Real Estate Sentiment Dictionary* are drawn in the eighth section.

4.2 Literature Review

Winson (2017) provides a good definition of studies on market sentiment: they "[…] analyze different sources of information to assess the prevailing attitude or mood of investors towards a given market or asset class, making qualitative judgements that are used to predict directionality." The notion underlying this definition is that decision-making processes are often not based purely on information about fundamentals, but are also influenced by further factors causing market movements. One current stream of research focuses on developing and testing ways to quantify sentiment extracted from textual data and accordingly evaluates the value added for traditional asset pricing models.

Over the last few years, academic research has increasingly confirmed the value of textual analysis to gain insights into market sentiment. The ongoing rise of the internet, accompanied by the digitization of all kinds of text corpora, and technical advances, have created various possibilities for a new field of sentiment analysis. Innovative methodologies have been developed, focusing on the evaluation of text data. There are two common streams of content analysis methods: machine learning and the dictionary-based approach.

The first was pioneered by computer scientists and mathematicians based on statistical techniques (Li, 2010). Algorithms such as Naïve Bayes classifiers or Support Vector Machines are trained with a pre-classified data set in a first step. The training data set can, for example, be annotated manually (Li, 2010) or by using an existing sentiment lexicon, as by Das and Chen (2007). In a second step, the algorithm applies the "learned" classification rules to annotate each text entity with one of the pre-defined sentiment categories. One of the earliest works by Antweiler and Frank (2004) conducted textual analysis with both the Naïve Bayes and the Support Vector Machine algorithm and found that bullishness indices extracted from stock message postings on *Yahoo! Finance* and *Raging Bull* are related to future stock trading volume.

Secondly, the dictionary-based approach is well-established in the literature. The methodology is based on word lists, in which each word is pre-annotated with a sentiment category. These word lists are often referred to as "sentiment dictionary" or "sentiment lexicon". In order to measure the sentiment of a text corpus, the researcher counts the occurrence of words from the pre-annotated word list, scaled by the total number of words in the text document. Tetlock (2007) popularized this methodology in the finance literature by demonstrating that high media pessimism extracted from *Wall Street Journal* newspaper articles predicts downward pressure on stock market prices. Furthermore, he found that unusually high or low values of his pessimism measure lead to higher trading volume. In 2008, Tetlock *et al.* again used the *Harvard IV-4* psychosocial dictionary to extract the fractions of negative words from financial news stories between 1980 and 2004. This paper confirms earlier findings of a negative relationship between media pessimism and stock prices, this time at the firm level. These studies were the starting point for future research. Applying the *Harvard IV-4 dictionary,* Engelberg (2008) found evidence of a significant linkage between *Dow Jones News Service* stories about firms' earnings announcements and subsequent asset prices. Following the same example, Frankel *et al.* (2010) quantified the linguistic tone of quarterly earnings conference call transcripts.

Another milestone in the evolution of the dictionary-based approach was the work of Loughran and McDonald in 2011. They investigated the notion that words have different meanings in different contexts – hence, a general dictionary might lead to misclassifications in a specific context. Loughran and McDonald (2011) analyzed 10-Ks from 1994 to 2008 and found that the application of the general *Harvard IV-4 dictionary* resulted in a misclassification of almost three-quarters of the words identified as negative. Consequently, they developed word lists which aimed to capture the meaning of words specifically in a financial context. Loughran and McDonald (2011) created six word lists, which are publicly available: negative, positive, uncertainty, litigious, strong modal, and weak modal annotated words. Over the following years, they improved and extended their word lists continuously. Similarly, Loughran and McDonald (2015) discovered that the use of Diction, a platform that enables tabulating words into pre-defined functional categories, so as to gauge sentiment in a financial context, is inappropriate. Many researchers, such as Doran *et al.* (2012), Engelberg *et al.* (2012), Price *et al.* (2012), Ferris *et al.* (2013), or Heston and Sinha (2017) adapted this notion and compared the applicability of different dictionaries in various contexts. Price *et al.* (2012), for example, found that using the finance-specific Loughran and McDonald (2011) sentiment dictionary leads to a better detection of the tone of relevant quarterly earnings conference calls, than using the general psychological *Harvard IV-4 dictionary*. Ferris *et al.* (2013) further confirmed these results by analyzing IPO prospectuses. Examining over 900,000 news stories among others, with the dictionary-based approach, Heston and Sinha (2017) extracted sentiment with the help of the Harvard psychological dictionary and the financial dictionary of Loughran and McDonald (2011). They found that the specialized financial dictionary is superior to the general dictionary for extracting sentiment that is relevant to future stock returns. This evidence from the literature further confirms the importance of an appropriate sentiment-annotated word list, either for each dictionary-based approach or for some machine learning approaches.

Some first attempts at applying the dictionary-based approach in the real estate literature include Doran *et al.* (2012), Walker (2014, 2016), Soo (2015), and Ruscheinsky *et al.* (2018). Doran *et al.* (2012) examined the tone of quarterly earnings conference calls from Real Estate Investment Trusts. Their results yield robust evidence of the predictive power of sentiment measures on contemporaneous stock price reactions. Walker (2014, 2016) twice tested the relationship between news media and the UK housing market. Applying the dictionary-based approach to housing-related news led to valuable sentiment measures. Likewise, Soo (2015) quantified the tone of housing news and found a leading linkage from sentiment on future house prices in the US. Ruscheinsky *et al.* (2018) contributed to this stream of literature by analyzing 125,000 news-media article headlines from four different source, namely *Bloomberg*, *The Financial Times*, *Forbes,* and *The Wall Street Journal.* In a vector autoregressive framework, they found significant evidence of a positive relationship between media-expressed sentiment and three- and four-month-ahead REIT market movements.

However, no research has been found so far that concentrates on the German real estate market. Accordingly, this paper pioneers textual analysis-based sentiment extraction for German real estate text corpora. As mentioned above, the first step is to choose an appropriate sentiment-annotated word list. Until now, there have been some first attempts to summarize presumably sentiment-bearing words. *German Polarity Clues* is a publicly available sentiment-annotated word list, developed by a semi-automatic translation approach from existing English resources into German (Waltinger, 2010). Similarly, the *SentiWS* dictionary uses translations from the General Inquirer by *Google Translate* and was subsequently revised manually (Remus *et al.*, 2010). Both pre-annotated sentiment word lists are constructed for a broad, general usage and relatively error-prone due to translation difficulties. The textual analysis literature generally agrees that discipline-specific dictionaries lead to fewer misclassified results, but no appropriate dictionary for an application in the German real estate market is available so far. Therefore, this paper provides a foundation by developing a *German Real Estate Sentiment Dictionary*.

4.3 Creation of the German Real Estate Sentiment Dictionary

Dictionary-based textual analysis stands or falls with the quality and relevance of the dictionary. A commonly used source for word classifications is the Harvard psychological dictionary, especially the *Harvard IV-4* negative word list (Tetlock, 2007; Engelberg, 2008; Tetlock *et al.*, 2008). Among others, Loughran and McDonald (2011) claim that discipline-specific dictionaries can reduce measurement errors. Each discipline has its own word meanings, which may not translate and apply effectively across different disciplines. As neither well-established general dictionaries like the *Harvard IV-4* word list for the German language, nor a real estate-specific word list are available, one objective of this paper is to develop a discipline-specific sentiment dictionary for the German real estate industry. This dictionary is intended to comprise real estate-related words with a clear positive or negative connotation in a real estate context.

By conducting an online survey among 1,686 respondents, relative objectivity in terms of the word classification can be achieved. The development of the *German Real Estate Sentiment Dictionary (GRESD)* is structured in the following three main steps:

Step 1: Creation of an extensive word list of real estate-related words, which are assumed to have a positive or negative tone in real estate contexts

Step 2: Classification of word list by survey participants into one of three categories: positive, neutral, or negative

Step 3: Evaluation of the survey and creation of the *German Real Estate Sentiment Dictionary*

4.3.1 Step 1: Creation of Word List

The basis of the survey is a list of real estate-related words, which presumably bear sentiment. Nouns, verbs, adjectives, adverbs, and prepositions are extracted from existing general German sentiment dictionaries, namely from the *German Polarity Clues* of Waltinger (2010) and the *SentiWS* of Remus *et al.* (2010). Furthermore, words that are likely to capture sentiment from the German real estate dictionary *Wörterbuch Immobilienwirtschaft* of Schulte *et al.* (2011) are included. Words with ambiguous meanings, swear words or colloquial vocabulary, as well as expressions with more than one word, are excluded. In order to ensure that each word is indeed used in a real estate context, the word list is verified by checking the occurrence of each word in the *Immobilien Zeitung* between 1995 and 2017. This ensures that only words constituting real estate jargon are included in the final *GRESD*. Next, the large collection of words is reduced to their basic forms (lemmas), in order to obtain a reasonable number of words, which is feasible for survey use. The result is a word list comprising 2,245 lemmas.

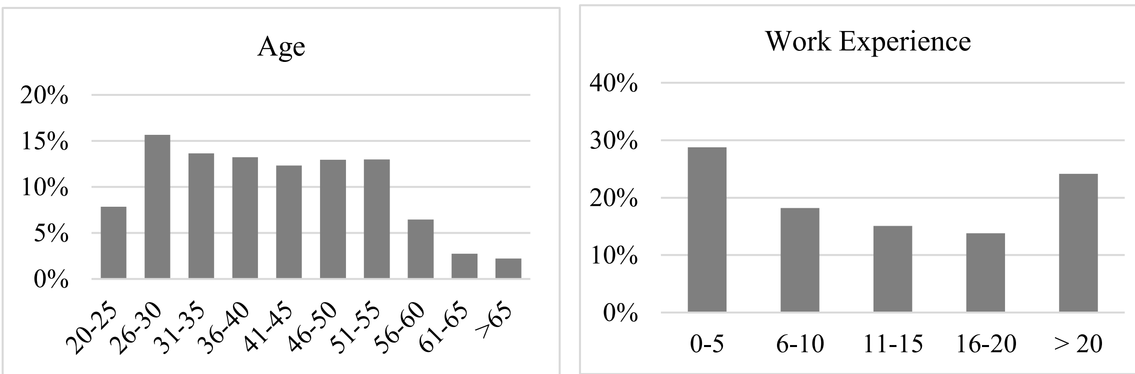
4.3.2 Step 2: Online Survey and Respondent Profiles

For the survey, an email-based questionnaire divided into two parts was developed. In order to determine whether the respondents constitute a representative sample of real estate professionals, the first part contains personal questions about the respondents' gender, age, work experience, company size, qualifications, etc. Closed questions with specified response options are employed to reduce the answering effort for respondents and to facilitate the questionnaire analysis (Krosnick and Presser, 2009).

In the second part of the survey, each participant is asked to classify 30 randomly selected words as positive, neutral or negative. To increase the response rate and to shorten the survey duration, the number of words was limited to 30 per respondent. The participants were asked to spend no longer than five seconds on each word classification in order to identify their spontaneous feeling when reading a specific word. The layout of the questionnaire is shown in Appendix 4.1. The survey was sent exclusively to German real estate professionals.

The results of the personal questions provide insights into whether the respondent sample of 1,686 respondents is representative or not. The respondents comprised 34% females and 66% males, with the relatively high proportion of male participants being inevitable, due to the distribution of human resources in the real estate industry in general. Respondents' age and work experience in years is shown in Exhibit 4.1.

Exhibit 4.1 | Respondents' age and work experience in years



Notes: The exhibit depicts the distribution of age and wok experience in years.

The age distribution comprises a very well-balanced sample, with the highest proportion of 16% in the age group 26-30 years. Overall, the age groups (from 26-55 years) are all represented, with more than 12% per group. Furthermore, the sample is dominated by respondents with work experience of either 0-5 years (29%) or over 20 years (24%). Other groups are distributed between 6 and 20 years of work experience. Furthermore, the respondents were asked to indicate their qualifications, their current position and the company size. 33% of the participants have a diploma (German "Diplom") followed by 23% with a master's and 14% with a bachelor's degree. The majority of participants works as employees without management or operational responsibility (41%) in a company with 51-250 staff members (27%). The participants were also asked in which field of the real estate industry they work. The highest proportions are employed in Asset-, Property- & Facility Management (13%) and Real Estate Development (11%). Overall, the respondents work in more than 20 different sectors of the industry. A detailed respondent profile description is shown in Appendix 4.2.

In summary, the respondent profiles represent a well-diversified sample of the German real estate industry, as no question delivered a biased distribution. Hence, the survey is representative for the entire German real estate industry and thus yields reliable sentiment-classification results.

4.3.3 Step 3: Development of the German Real Estate Sentiment Dictionary

In the course of sentiment classification, each word was classified by at least 21 different respondents as positive, neutral or negative. This means, that all words were classified 21 times, some words even 22 times, but round 22 was not completed for all words. As a threshold for a word being included in the *German Real Estate Sentiment Dictionary*, it has to be classified as positive or negative by more than 50% of the respondents. Neutral words are not included in the *GRESD,* as they do not have any explanatory power for sentiment analysis.

The lemma word list is subsequently expanded by their inflections. Finally, the *GRESD* comprises 8,330 (59%) negative and 5,807 (41%) positive words, which results in a list of 14,137 sentiment-bearing German words. Exhibit 4.2 provides a comprehensive overview of the dictionary composition.

Exhibit 4.2 | The German Real Estate Sentiment Dictionary

	Lemmas			Inflections included		
	Negative lemmas	**Positive lemmas**	**Total**	**Negative words**	**Positive words**	**Total**
Nouns	695	341	1,036 *(55%)*	2,158	1,024	3,182 *(23%)*
Verbs	157	77	234 *(12%)*	1,569	744	2,313 *(16%)*
Adjectives	325	279	604 *(32%)*	4,597	4,034	8,631 *(61%)*
Adverbs	4	5	9 *(0%)*	5	5	10 *(0%)*
Prepositions	1	0	1 *(0%)*	1	0	1 *(0%)*
Total	1,182 *(63%)*	702 *(37%)*	1,884	8,330 *(59%)*	5,807 *(41%)*	14,137

Notes: The *GRESD* comprises 1,182 (63%) negative and 702 (37%) positive lemmas. Including all inflections for each word, the dictionary results in 14,137 words with 8,330 (59%) negative and 5,807 (41%) positive ones. The numbers in italics state the breakdown of the total number of words.

In comparison to the *GRESD*, the *Harvard IV-4* word list consists of 2,291 (54%) negative and 1,915 (46%) positive lemmas and the discipline-specific dictionary of Loughran and McDonald

(2011), of 2,337 (87%) negative and 353 (13%) positive words. Hence, regarding these numbers, the *German Real Estate Sentiment Dictionary* represents a comprehensive word list with a well-balanced proportion of negative and positive words, thus preventing biased results. The *German Real Estate Sentiment Dictionary* is freely accessible online at www.irebs.de (Link will be activated once the paper is accepted for publication).

4.4 Data

The data used in this paper consists of two different sets. The first is a text corpus comprising newspaper articles and the second, a direct real estate price index, as well as macroeconomic variables. The sample period extends from February 2007 to October 2017. The data frequency is monthly, which results in 129 observations in total.

4.4.1 Text Corpus

The text data includes all newspaper articles of the *Immobilien Zeitung (IZ)* published between 2007 and 2017. There are several reasons for choosing the news provider *Immobilien Zeitung* to quantify German real estate-related sentiment. Firstly, the *IZ* is one of the leading and most well-established newspapers in the German real estate industry. Secondly, it covers current events in the real estate market and provides background information, market data, as well as people and company news. Thirdly, in terms of data availability, electronic texts are available from 1995 onwards. The sample covers 125,462 print and online articles with about 27 million words in total. With an average article length of 222 words, the articles of the *Immobilien Zeitung* are relatively short compared to other international newspapers like the *New York Times* (1,021 words per article) or the *Huffington Post* (641 words per article) (Newswhip, 2017). Therefore, the *IZ* constitutes a real estate news portal, which provides compact industry-related news.

All texts are tokenized to decompose text into single words and punctuation was removed. Pictures, graphs, tables, English articles, and editorial shortcuts were excluded. The average number of articles per month was about 345 in 1995 and increased up to 1,131 in 2017.

4.4.2 Real Estate and Macroeconomic Data

To replicate the German residential real estate market, the *IMX Immobilienindex* is used. This is a real estate price index based on over 12 million real estate residential offers on Germany's most popular home search website *ImmobilienScout24*. This website is shown to be the market leader amongst residential real estate portals in Germany as 63% of all home-seekers use this website in order to find their new home (comScore, 2016). In this paper, the price index for new apartments is employed as a measure of the direct German real estate market. The *IMX* has been publicly available since the beginning of March 2007 (end of February 2007) on a monthly basis.

Based on previous empirical evidence from the German and US housing market, this paper includes macroeconomic control variables thought to influence real estate returns (Cieleback, 2012; Walker, 2016; Freybote and Seagraves, 2017). Given that several housing studies highlight the impact of labor market variables on housing demand (Nakajima, 2011; Soo, 2015), the number of unemployed people (*UNEMP*) and the average wages and salaries in the overall economy (*WAGES*) are incorporated. To replicate the current economic situation, the industry turnover of capital goods (*INDTURN*) is included. Furthermore, building permits (*BUILDPER*) and construction turnover (*CONSTURN*) are considered as two proxies for the housing supply. In addition, the home loan interest rate (*INT*), which has been shown to influence housing demand and prices (Mayer and Sinai, 2009; Taylor, 2013), is used.

Exhibit 4.3 | Descriptive statistics

	Mean	Median	Max	Min	SD	Datasource
IMX	0.45%	0.40%	1.78%	-0.76%	0.35%	ImmobilienScout24
UNEMP	-0.38%	-1.20%	12.48%	-5.07%	3.66%	Destatis
WAGES	2.47%	0.05%	55.49%	-34.07%	20.66%	Datastream
INDTURN	1.16%	-0.45%	37.01%	-28.75%	14.61%	Datastream
BUILDPER	0.74%	0.02%	34.95%	-34.87%	11.36%	Destatis
CONSTURN	3.51%	4.92%	92.99%	-63.78%	21.09%	Destatis
INT	-0.61%	-0.72%	1.06%	-2.74%	0.65%	Deutsche Bundesbank

Notes: This table reports descriptive statistics of monthly variables between 2007 and 2017. *IMX* is the growth rate of the German real estate price index for new apartments. *UNEMP* is the growth rate of the number of unemployed people. *WAGES* is the growth rate of wages and salary for the overall economy. *INDTURN* is the growth rate of whole industry turnover. *BUILDPER* is the growth rate of construction permits. *CONSTURN* is the growth rate of construction turnover and *INT* is the growth rate of residential loan interest rate. The sources of the variables are named accordingly in the last column of the table.

Exhibit 4.3 provides descriptive statistics of *IMX* returns and the control variables. Mean, median, maximum, minimum, and standard deviation of growth rates are reported in decimal form. The residential real estate market for new apartments averages a monthly growth rate of 0.45%. The monthly growth rate of unemployed people (*UNEMP*) is very stable, with a low standard deviation of 3.66% due to seasonal employment, whereas the salary (*WAGES*) and turnover of the construction industry (*INDTURN*) show a higher standard deviation of about 20%. As expected, the average growth rate of home loan interest rates (*INT*) is negative at -0.61%, because of decreasing interest rates over the last years. The *IMX* prices, as well as all control variables, were transformed into growth rates to address non-stationarity issues.

4.5 Methodology

4.5.1 Dictionary-based Approach

In order to extract sentiment from newspaper articles, the dictionary-based approach is applied. This methodology belongs to the "bag-of-words" approaches (Loughran and McDonald, 2016), because it separates each word from a text corpus and treats it as an individual entity. Consequently, the order and co-occurrence of words are ignored. With a pre-annotated sentiment dictionary, the number of words belonging to one of the sentiment categories can be counted. Referring to Loughran and McDonald (2016), this methodology has three main advantages. First, the subjectivity of researcher decisions is largely avoided, because the classification is based solely on the dictionary. Second, the methodology is easily scalable with an appropriate computer program and with publicly available dictionaries, and thirdly, the analysis process is more straightforward to replicate than most others.

To meet the crucial challenge of negation in textual analysis, this is treated similarly as in Soo (2015), by switching the sentiment if one of the predetermined 20 negation terms is present in a window of five words before the occurrence of the sentiment-bearing word. The number of positive and negative words is added up separately for each text entity. As the dataset is quite large, RapidMiner is utilized to conduct the counting task. RapidMiner is a software platform for data science applications, such as data preparation, machine learning, deep learning, text mining, or predictive analysis (https://rapidminer.com/). The final process constructed to perform the dictionary-based approach on textual data is replicable and can be seen in Appendix 4.3.

4.5.2 Real Estate Sentiment Measures

The results from the dictionary-based approach give the amount of positive and negative words for each text entity. Depending on the time period of interest, the sentiment measures can be aggregated at different frequencies. This is a major advantage, compared to traditional sentiment measures which are usually based on surveys, because sentiment indicators can be constructed for any desired frequency – quarterly, monthly, weekly, daily, or even hourly – the only restrictive factor being the number of text entities available for analysis. Due to the limited availability of the *IMX* and the macroeconomic variables, a monthly analysis is chosen. Hence, the sentiment extracted from each text entity is aggregated to a monthly level. Two different sentiment measures are introduced; one focusing on negative text entities and one on positive texts, scaled by the total number of newspaper articles taken into account for the time period. The *Negative Indicator* is calculated by:

$$NI_t\_D = \frac{\sum_1^I negative\ text\ entity_{i,t}}{\sum total\ number\ of\ text\ entities_t} \tag{10}$$

and the *Positive Indicator* accordingly as:

$$PI_t\_D = \frac{\sum_1^I positive\ text\ entity_{i,t}}{\sum total\ number\ of\ text\ entities_t} \tag{11}$$

where i is a text entity classified as negative or positive and t is the period in which all text entities must be published in order to be taken into account. D relates to the dictionary threshold used to assign words to positive or negative sentiment categories. As the *GRESD* is constructed by conducting a survey, each word has a percentage score, showing how often it was classified positive or negative in relation to the total amount of responses. For example, the word "Verlust" (translation: loss) was classified negative by 18 and neutral by 4 respondents, resulting in a negativity score of 82%. "Flexibilität" (translation: flexibility), for example, was classified positive by 17, neutral by 4, and negative by 1 respondent(s), which yields a positivity score of 77%. This sentiment scoring for each word enables constructing sentiment dictionaries with different thresholds. Meaning the *GRESD* with the threshold of 50% (*GRESD_50*) includes all words which are classified positive or negative by at least 50% of the respondents, *GRESD_60* by at least 60% and so on.

4.5.3 Vector Autoregressive Framework

To test the relationship between different sentiment measures based on the developed *German Real Estate Sentiment Dictionary* and the German residential real estate market, a vector autoregressive framework is deployed. Before conducting any regression analysis, all variables of the vector autoregressive model are tested using an *Augmented Dickey-Fuller Test* to check for the existence of a unit root. Whenever the required stationarity was rejected, variables are differenced or used as growth rates to ensure statistical appropriateness.

A vector autoregressive regression (VAR) is able to capture the dynamic relationship between endogenous variables and is flexible and compact in expressing the notation. In its simplest form, it just contains two variables, y_{1t} and y_{2t}, depending on different combinations of the previous k values of each other and error terms, the so-called bivariate VAR:

$$y_{1t} = \beta_{10} + \beta_{11}\, y_{1t-1} + \cdots + \beta_{1k}\, y_{1t-k} + \alpha_{11}\, y_{2t-1} + \cdots + \alpha_{1k}\, y_{2t-k} + u_{1t} \tag{12}$$

$$y_{2t} = \beta_{20} + \beta_{21}\, y_{2t-1} + \cdots + \beta_{2k}\, y_{2t-k} + \alpha_{21}\, y_{1t-1} + \cdots + \alpha_{2k}\, y_{1t-k} + u_{2t} \tag{13}$$

where u_{it} is a white noise disturbance term with $E(u_{it}) = 0$, ($i = 1, 2$), $E(u_{1t}, u_{2t}) = 0$ (Brooks and Tsolacos, 2010).

As this paper investigates the value of media sentiment on direct real estate market movements, further influencing factors have to be included as well. Hence, the final model denoted in a short matrix notation includes X as a matrix of further exogenous model variables and B as a matrix of corresponding coefficients:

$$y_t = A_0 + A_1 y_{t-1} + \cdots + A_k y_{t-k} + BX + u_t \tag{14}$$

Vector autoregressive models are highly sensitive to the lag length. Hence, determining the optimal lag length is crucial. One method recommended by Brooks and Tsolacos (2010) is to use information criteria such as Aikaike's information criterion (AIC), Schwarz's Bayesian information criterion (SC), or the Hannan-Quinn information criterion (HQIC). These criteria contain two factors with different characteristics, one factor is a function of the residual sum of squares (RSS) and the other factor penalizes for the loss of degrees of freedom by adding further parameters to the model. Hence, including an additional lag will cause the RSS to fall, but at the same time, the penalty term will increase. From this scenario, it follows that the lag length that minimizes the value of the information criteria should be chosen.

To test whether changes in y_2 – here the created sentiment measures – cause changes in y_1 – here the direct real estate market – and vice versa, the Granger Causality test is applied to each model. Furthermore, the *Breusch-Godfrey Lagrange Multiplier Test* is conducted for each model, so as to ensure that the residual series are not serially correlated.

4.6 Results

4.6.1 Relationship between Sentiment Measures and the IMX Price Index

Based on economic theory, a vector autoregressive model is derived to explain the residential real estate market returns between 2007 and 2017. The regressions are conducted on a monthly basis and control for the same set of macroeconomic variables each time, namely unemployment, building permits, construction turnover, industry turnover, wages, and home loan interest rate. The *IMX* returns and the sentiment measures are included in the vector autoregressive framework as endogenous variables and the controls as exogenous variables. All models are stable regarding common robustness tests and indicate an optimal lag length of three. As shown in Exhibit 4.4, the analysis starts with *Negative* and *Positive Indicators*, which are calculated by applying the *German Real Estate Sentiment Dictionary* with the threshold of 70%, in order to analyze the complete sample of newspaper articles (headlines and full text).

Exhibit 4.4 | VAR estimation results with *Negative* and *Positive Indicator*

	IMX					
	Model 1		Model 2		Comparison Model	
	Negative Indicator		*Positive Indicator*		-	
IMX (-1)	0.485	***	0.443	***	0.462	***
	[5.87851]		[5.19546]		[5.51991]	
IMX (-2)	-0.170	*	-0.137		-0.160	*
	[-1.78406]		[-1.39351]		[-1.66332]	
IMX (-3)	-0.210	**	-0.164	*	-0.204	**
	[-2.53134]		[-1.84989]		[-2.42480]	
Sentiment (-1)	-0.039	**	-0.013			
	[-2.53628]		[-1.25516]			
Sentiment (-2)	0.001		-0.014			
	[0.04977]		[-1.25201]			
Sentiment (-3)	0.020		-0.003			
	[1.24851]		[-0.26684]			
Constant	-0.004	**	-0.004	**	-0.004	**
	[-2.39240]		[-2.49936]		[-2.51878]	
Macroeconomic Controls	YES		YES		YES	
R^2	0.571		0.542		0.531	
Adj. R^2	0.521		0.488		0.490	
Log likelihood	580.948		576.843		575.451	
Akaike AIC	-9.071		-9.005		-9.031	
Schwarz SC	-8.754		-8.689		-8.782	
Granger Causality						
Sentiment indicator	10.205	**	2.501			
IMX	0.072		6.863	*		

Notes: This table reports results for the estimated VAR models with 3 lags, monthly *IMX* returns, and news-based sentiment as endogenous variables. The set of macroeconomic control variables includes the following: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates with no lag, industry turnover growth rate with a fourth lag, wages with a first lag, and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

Model 1, hereinafter referred to as *Base Model*, shows the dynamic relationship between the *IMX* price index returns and the *Negative Indicator (NI)*. The *NI_70* exerts a statistically significant influence on one-month-ahead *IMX* returns. The first lag of the *Negative Indicator* shows the expected negative sign and is significant at the 5% level. This result is backed by the associated Granger Causality, which demonstrates that the sentiment measure *NI_70* influences the *IMX* movements beyond all control variables and past returns of the *IMX* itself. One explanation for the relatively quick impact of sentiment on prices might be that less informed investors are more prone to sentiment than informed investors (Garcia, 2013). Most residential investors are supposedly less informed than institutional investors, for example. Furthermore, the amount of transactions in the residential market is higher than in other real estate markets, whereas the transaction volumes are lower. Hence, sentiment changes might be incorporated into housing prices faster.

Compared to the VAR with no sentiment measure, the *Comparison Model*, the inclusion of the *Negative Indicator* enhances the R-squared by 7.5% from 53.1% to 57.1% and the adjusted R-squared by 6.3% from 49% to 52.1%. As the Log likelihood increases and the information criteria decreases, the whole model improves in terms of goodness of fit by including the sentiment measure. An interesting aspect worth mentioning is that the *IMX* does not Granger-cause the *Negative Indicator*. Hence, this causality is unidirectional. However, the *Positive Indicator*, based solely on text corpora classified as positive, does not show any significance in explaining future *IMX* movements. Moreover, the *PI_70* presents negative coefficients. This entails the Granger Causality of the *PI_70* also not being statistically significant.

These findings indicate a negativity bias of German real estate market participants. The bias refers to the notion that humans accord greater relevance to negative entities and results than to positive ones, even when both are of the same magnitude. This phenomenon has already been discovered and discussed in the psychological literature (Rozin and Royzman, 2001). Furthermore, Tetlock (2007) developed this idea by creating pessimism indicators only. He found a significant relationship between media pessimism and subsequent stock market prices.

4.6.2 Importance of Sentiment Dictionary Compilation

Several studies emphasize the importance of choosing an appropriate sentiment dictionary, because it is the basis for at least the dictionary-based approach, and in some cases for machine learning approaches as well. Due to the classification of the world list by survey participants, it is possible to test different design scopes for the *German Real Estate Sentiment Dictionary*. As the survey yielded percentage amounts for each word, referring to how often it was classified by the respondents as positive, neutral and negative, it is possible to construct dictionaries with different thresholds. For example, the adjective “wertvoll” (valuable) was classified positive by 19 respondents and neutral by 3, resulting in a positivity score of 86%. In Exhibit 4.5, *Negative Indicators* constructed with different dictionary manifestations are included in the VAR framework.

Exhibit 4.5 | VAR estimation results for sentiment measures based on dictionaries with different thresholds

	IMX				
	Model 3	Model 4	*Base Model*	Model 5	Model 6
	NI_50	*NI_60*	*NI_70*	*NI_80*	*NI_90*
IMX (-1)	0.478 ***	0.488 ***	0.485 ***	0.475 ***	0.461 ***
	[5.68780]	[5.83799]	[5.87852]	[5.65732]	[5.48456]
IMX (-2)	-0.189 *	-0.180 *	-0.170 *	-0.171 *	-0.150
	[-1.92399]	[-1.85897]	[-1.78403]	[-1.76201]	[-1.54554]
IMX (-3)	-0.200 **	-0.211 **	-0.210 **	-0.193 **	-0.205 **
	[-2.33829]	[-2.50007]	[-2.53136]	[-2.27758]	[-2.38338]
Sentiment (-1)	-0.021	-0.031 **	-0.039 **	-0.020	-0.009
	[-1.34626]	[-2.11146]	[-2.53629]	[-1.20633]	[-0.42879]
Sentiment (-2)	-0.013	-0.023	0.001	-0.006	0.009
	[-0.78515]	[-1.37487]	[0.04977]	[-0.30632]	[0.42333]
Sentiment (-3)	0.014	0.011	0.020	0.017	0.029
	[0.91114]	[0.74186]	[1.24850]	[1.05069]	[1.44340]
Constant	-0.004 **	-0.004 **	-0.004 **	-0.004 **	-0.004 **
	[-2.53645]	[-2.46744]	[-2.39240]	[-2.44515]	[-2.51074]
Macroeconomic controls	YES	YES	YES	YES	YES
R^2	0.548	0.559	0.571	0.545	0.543
Adj. R^2	0.495	0.508	0.521	0.491	0.489
Log likelihood	577.683	579.283	580.948	577.270	576.989
Akaike AIC	-9.019	-9.045	-9.071	-9.012	-9.008
Schwarz SC	-8.702	-8.728	-8.754	-8.696	-8.691
Granger Causality					
Sentiment indicator	4.035	7.019 *	10.205 **	3.278	2.765
IMX	2.481	1.109	0.072	3.122	8.725
Number of words in dictionary	14,137	12,851	10,563	7,363	3,219

Notes: This table reports results for the estimated VAR models with 3 lags, monthly *IMX* returns, and news-based sentiment, using different thresholds of the *GRESD* as endogenous variables. The set of macroeconomic control variables includes: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates with no lag, industry turnover growth rate with a fourth lag, wages with a first and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

The results yield a significant predictability of the *Negative Indicator_60* and the *Negative Indicator_70* on future residential real estate market movements. As mentioned above in the *Base Model*, the *NI_70* Granger-causes *IMX* returns at the 5% significance level. The *NI_60* presents similar findings, with a Granger Causality at the 10% significance level. Equally to the *Base Model*, the first lag of the *NI_60* is statistically significant and has the expected negative sign. However, no other negative indicators based on the thresholds of 50%, 80%, and 90% (Models 3, 5, 6) show any predictive power with respect to subsequent *IMX* returns. This seems reasonable, especially for the threshold of 90% classification accordance, as in this case, only 3,219 words are left in the *GRESD*. This amount might be too low to identify enough sentiment bearing words. Looking at the other extreme, a possible explanation of the non-existent relationship between the *NI_50* and *IMX* returns is that too many words are used for the sentiment-extraction process, which might have ambiguous meanings. The *NI_70* yields the strongest results in terms of both statistical significance and the overall goodness of fit of the model. By containing different classification accordance levels, the *GRESD* empowers future researchers to individually determine which threshold is suitable for investigating their specific research questions.

4.6.3 Investigating Different Parts of the Newspaper Article

Besides the choice of an appropriate dictionary, another decision to make is which text data to analyze. This decision is equally important in order to capture relevant market sentiment. First, the data source and hence, the news quality is decisive. Second, it must be decided which parts of a newspaper and, which parts of a particular newspaper article should be selected.

Different notions about which part of newspaper articles should be investigated can be found in the literature. Strapparava and Mihalcea (2008) describe headlines of news articles as especially suitable for textual sentiment analysis as they are short, written to attract reader attention, and are often provocative. As the dataset of this present paper is clearly-structured, it is possible to distinguish and make comparisons between the predictive power of the sentiment measures extracted from different parts of a newspaper article. The question arises, whether the analysis of the headline alone is already enough to capture market sentiment.

Exhibit 4.6 shows the results of the *Negative Indicators* based on the *GRESD_60* and *GRESD_70* analyzing headlines alone (H) in the first step, and the full text of all newspaper articles in the second step (HT). The results indicate that the analysis of newspaper headlines alone is indeed already sufficient to determine predictive power with respect to future residential real estate market movements. In both Models 7 and 8, the *Negative Indicator* Granger-causes *IMX* price changes significantly at the 5% level, even when controlling for various macroeconomic fundamentals. As comparison, both models analyzing the headlines and full text of all newspaper articles with the *GRESD_60* and *_70* are reported again. As mentioned earlier, they both have statistically significant explanatory power with respect to the *IMX* return changes. The results further indicate the robustness of the relationship found.

Exhibit 4.6 | VAR estimation results for sentiment measures based on different parts of a newspaper article

	IMX			
	NI_ 60		*NI_70*	
	Model 7	Model 4	Model 8	*Base Model*
	H	*HT*	*H*	*HT*
IMX (-1)	0.477 ***	0.488 ***	0.475 ***	0.485 ***
	[5.74123]	[5.83799]	[5.71930]	[5.87852]
IMX (-2)	-0.164 *	-0.180 *	-0.161 *	-0.170 *
	[-1.72773]	[-1.85897]	[-1.69659]	[-1.78403]
IMX (-3)	-0.190 **	-0.211 **	-0.181 **	-0.210 **
	[-2.28047]	[-2.50007]	[-2.17973]	[-2.53136]
Sentiment (-1)	-0.049 *	-0.031 **	-0.076 **	-0.039 **
	[-1.86226]	[-2.11146]	[-2.31578]	[-2.53629]
Sentiment (-2)	-0.017	-0.023	-0.012	0.001
	[-0.58718]	[-1.37487]	[-0.36106]	[0.04977]
Sentiment (-3)	0.031	0.011	0.030	0.020
	[1.18864]	[0.74186]	[0.91364]	[1.24850]
Constant	-0.004 **	-0.004 **	-0.004 **	-0.004 **
	[-2.47530]	[-2.46744]	[-2.51260]	[-2.39240]
Macroeconomic variables	YES	YES	YES	YES
R^2	0.564	0.559	0.570	0.571
Adj. R^2	0.513	0.508	0.520	0.521
Log likelihood	579.928	579.283	580.832	580.948
Akaike AIC	-9.055	-9.045	-9.069	-9.071
Schwarz SC	-8.738	-8.728	-8.753	-8.754
Granger Causality				
Sentiment indicator	8.242 **	7.019 *	9.980 **	10.205 **
IMX	3.624	1.109	4.333	0.072

Notes: This table reports results for the estimated VAR models with 3 lags, monthly *IMX* returns, and sentiment measures based on different parts of the newspaper article. The set of macroeconomic control variables includes: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates with no lag, industry turnover growth rate with a fourth lag, wages with a first and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

Overall, both variations lead to statistically significant results, but it is worth mentioning the trade-off concerning which parts of the newspaper article to use. On the one hand, as headlines are short and generally summarize the article, it should be easy to capture the sentiment. On the other hand, the headline might have not enough words to identify any sentiment. Regarding complete texts, there are normally enough words to capture sentiment, but articles often contain conflicting views and balance various reasons and factors in the end. Therefore, the dictionary-based approach might not be able to capture this assessment adequately.

4.6.4 Comparison to General German Sentiment Dictionaries

Following the findings of Loughran and McDonald (2011), Price *et al.* (2012), and Heston and Sinha (2017), a comparison with two generic German linguistic sentiment dictionaries is performed in order to determine whether the created discipline-specific *German Real Estate Sentiment Dictionary* leads to superior sentiment measures.

Positive and *Negative Indicators* are constructed employing two general dictionaries, namely *SentiWS* and *German Polarity Clues* (*GPC*). The dictionary-based approach is applied to the complete newspaper articles accordingly. The descriptive statistics of the developed sentiment measures for each dictionary and their correlations are shown in Exhibit 4.7. The *Negative Indicators* of all three variations are around 10% on average, whereas the *Positive Indicator* lies between 46% and 82%. However, standard deviations are similarly low for all sentiment measures. As expected, *Positive* and *Negative Indicators* are in all three cases negatively correlated with a similar altitude.

Exhibit 4.7 | Descriptive statistics of different sentiment measures

	NI_70	*PI_70*	*GPC_NI*	*GPC_PI*	*SENTIWS_NI*	*SENTIWS_PI*
Mean	10.63%	46.15%	16.16%	73.64%	6.96%	82.24%
Median	10.35%	46.46%	16.09%	73.55%	6.67%	82.67%
Max	16.65%	56.30%	22.59%	80.07%	12.33%	89.39%
Min	5.64%	36.03%	11.11%	66.34%	3.83%	73.58%
Std. Dev.	2.48%	4.50%	1.89%	2.52%	1.61%	3.22%
Number of words in dictionary	6,282	4,281	19,962	17,627	10,155	10,103
Correlations						
NI_70	1.00					
PI_70	-0.51	1.00				
GPC_NI	0.60	-0.45	1.00			
GPC_PI	-0.53	0.76	-0.78	1.00		
SENTIWS_NI	0.78	-0.61	0.67	-0.68	1.00	
SENTIWS_PI	-0.57	0.81	-0.58	0.84	-0.77	1.00

Notes: This table provides descriptive statistics and correlations for the sample between 2007 and 2017. The number of words indicates how many positive or negative words are included in each dictionary.

Exhibit 4.8 | Comparison with general German sentiment dictionaries

	IMX					
	Model 1	Model 2	Model 9	Model 10	Model 11	Model 12
	GRESD	*GRESD*	*SentiWS*	*SentiWS*	*GPC*	*GPC*
	NI_70	*PI_70*	*Negative Indicator*	*Positive Indicator*	*Negative Indicator*	*Positive Indicator*
IMX (-1)	0.485 ***	0.443 ***	0.472 ***	0.488 ***	0.477 ***	0.482 ***
	[5.87851]	[5.19546]	[5.56973]	[5.74793]	[5.59574]	[5.66972]
IMX (-2)	-0.170 *	-0.137	-0.169 *	-0.197 **	-0.179 *	-0.186 *
	[-1.78406]	[-1.39351]	[-1.72919]	[-2.02228]	[-1.81725]	[-1.88276]
IMX (-3)	-0.210 **	-0.164 *	-0.205 **	-0.197 **	-0.196 **	-0.201 **
	[-2.53134]	[-1.84989]	[-2.41070]	[-2.33441]	[-2.29030]	[-2.34859]
Sentiment (-1)	-0.039 **	-0.013	-0.013	0.013	-0.011	0.015
	[-2.53628]	[-1.25516]	[-0.62667]	[1.09638]	[-0.91067]	[1.34956]
Sentiment (-2)	0.001	-0.014	-0.007	0.000	0.005	0.005
	[0.04977]	[-1.25201]	[-0.27041]	[0.01292]	[0.37384]	[0.37316]
Sentiment (-3)	0.020	-0.003	-0.019	0.015	0.000	0.010
	[1.24851]	[-0.26684]	[-0.93553]	[1.30698]	[-0.00401]	[0.92906]
Constant	-0.004 **	-0.004 **	-0.004 **	-0.004 **	-0.004 **	-0.004 **
	[-2.39240]	[-2.49936]	[-2.44060]	[-2.47865]	[-2.43835]	[-2.42567]
Macroeconomic Variables	YES	YES	YES	YES	YES	YES
R^2	0.571	0.542	0.537	0.549	0.538	0.542
Adj. R^2	0.521	0.488	0.483	0.496	0.484	0.488
Log likelihood	580.948	576.843	576.261	577.826	576.406	576.841
Akaike AIC	-9.071	-9.005	-8.996	-9.021	-8.999	-9.005
Schwarz SC	-8.754	-8.689	-8.679	-8.704	-8.682	-8.689
Granger Causality						
Sentiment indicator	10.205 **	2.501	1.447	4.299	1.710	2.496
IMX	0.072	6.863 *	0.846	3.655	0.741	2.203

Notes: This table reports results for the estimated VAR models with 3 lags, monthly *IMX* returns, and news-based sentiment as endogenous variables. The set of macroeconomic control variables includes: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates with no lag, industry turnover growth rate with a fourth lag, wages with a first and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

The VAR results in Exhibit 4.8 indicate no significant relationship between the sentiment measures based on the general dictionaries *SentiWS* (Models 9 and 10) or *GPC* (Models 11 and 12) and future residential real estate market returns. The negative and positive sentiment indicators do not yield any statistically significant coefficients, neither sentiment measures based on *SentiWS,* nor on *GPC* Granger-cause *IMX* returns. These findings support the argument of Loughran and McDonald (2011) among others, who state that a domain-specific dictionary is much more powerful in detecting sentiment. This confirms the quality and appropriateness of the *German Real Estate Sentiment Dictionary*.

4.6.5 Out-of-sample Forecasting

By comparing the forecasting from alternative models, the aim is to determine whether sentiment-augmented models achieve better results than models without any sentiment. Researchers agree that forecasting methods should be assessed using out-of-sample rather than in-sample tests, because in-sample errors are likely to understate forecasting errors (Tashman, 2000). The number of forecasting periods should not exceed the number of estimation periods. A sensible approach evaluating the forecasting accuracy is not to use all the observations in the estimation period but rather to hold some back. Hence, in this paper, the estimation period is defined from June 2007 to December 2016 and the forecasting period is from January 2017 to October 2017.

In order to compare the forecasting accuracy, this paper follows Brooks and Tsolacos (2010) and focuses on the forecast error $\hat{e}_{t+n,t}$, defined as the difference between the actual value of real estate returns (A_{t+n}) and the value of the forecast ($F_{t+n,t}$). This analysis concentrates on the variance-based forecasting error, namely the Root Mean Squared Error (RMSE), which is measured on the same scale as the data. Furthermore, the Theil's U1 coefficient constitutes an appropriate scalar for comparing forecasting accuracy of two different models (Theil, 1966, 1971). This coefficient ranges between zero and one, with coefficients closer to zero represent better predictions.

Exhibit 4.9 | Forecasting results

Model	Macro	Sentiment	Adj. *R2*	RMSE	RMSE Reduction	MSE	U1 Theil	U1 Theil Reduction
Model 7	x	-	0.486	0.00150		2.253E-06	0.110	
	x	*NI_60 H*	0.503	0.00119	20.5%	1.426E-06	0.088	19.8%
Model 4	x	-	0.486	0.00150		2.253E-06	0.110	
	x	*NI_60 HT*	0.497	0.00117	22.2%	1.364E-06	0.087	21.3%
Model 8	x	-	0.486	0.00150		2.253E-06	0.110	
	x	*NI_70 H*	0.508	0.00104	30.5%	1.088E-06	0.077	29.9%
Base Model	x	-	0.486	0.00150		2.253E-06	0.110	
	x	*NI_70 HT*	0.507	0.00098	34.7%	9.604E-07	0.073	33.3%

Notes: The reduction of the RMSE and of the U1 Theil coefficient is always measured in relation to the model without any sentiment measure. A positive value stands for an improvement in forecasting in comparison to the non-sentiment-model.

Exhibit 4.9 reports the dynamic forecasting accuracy results, namely the RMSE, MSE, U1 Theil statistic, and the RMSE and U1 Theil reduction of sentiment augmented models against models which include only macroeconomic control variables. The adjusted *R*-squared serves as a goodness of fit measure for the estimated model from 2007 to 2016 for all VAR models which have already been provided in Exhibit 4.6.

The most important finding is that models enriched with a sentiment measure have lower RMSE and U1 coefficients and thereby score better in terms of forecasting accuracy than the models without sentiment. For *NI_70* models, the RMSE reduction ranges between 30% and 35%, whereas for *NI_60* models, it lies between 20% and 22%. Theil's U1 coefficients of the sentiment-augmented models indicate that return forecasts for real estate residential prices come very close to their actual values, as U1 ranges from 0.07 to 0.11. The best forecasting accuracy measured by the U1 coefficient is achieved by the model that is augmented by *NI_70 HT*.

In the wake of RMSE and U1 reduction, it shows that for *NI_60* and *NI_70* the sentiment measures extracted from full texts (HT) show better forecasting accuracies. All in all, the inclusion of any kind of news-based sentiment measures apparently reduces forecasting errors for the residential real estate market in Germany. These results confirm that sentiment extracted from real estate-related newspaper articles contains relevant information which helps to explain and forecast residential real estate return movements.

4.6.6 Robustness

The results section already tests, in a variety of ways, the robustness of the relationship between the dictionary-derived news-based sentiment measures and the German residential real estate market. Nevertheless, this paper aims at conducting some final robustness tests in the following analysis. As the vector autoregressive framework is highly sensitive to the lag-length specification, the *Base Model* is run again with varying lag lengths. From Panel A in Exhibit 4.10 it can be seen that the first lag of the *NI_70* is robust at least at the 5% significance level, regardless of the total number of lags included. Models between one and five lags do result in a significant Granger Causality of the *Negative Indicator* on the *IMX*. All models are run with the same set of control variables as introduced in the *Base Model*. Choosing a lag length of six still presents a significant first lag of the *NI_70*, but jointly, the Granger Causality becomes insignificant for the first time.

Winsorizing the *IMX* shows whether or not the results are dependent on the extreme values of the times series. In a first step, the default is set to the 1% and 99% quantiles of the *IMX* and in a second step, to the 5% and 95% quantiles. Panel B shows that the results still hold. The *NI_70* significantly explains the winsorized *IMX* returns beyond their own past values and even when controlling for the set of macroeconomic variables. The Granger Causality is significant for both variations at the 5% significance level.

Exhibit 4.10 | Robustness tests

Panel A: Robustness of Granger Causality with Varying Lag Length								
	NI_70 $_{t-1}$	*NI_70* $_{t-2}$	*NI_70* $_{t-3}$	*NI_70* $_{t-4}$	*NI_70* $_{t-5}$	*NI_70* $_{t-6}$	Macro	$\chi 2$
Model with 1 lag	-0.035**						x	6.246**
Model with 2 lags	-0.042***	-0.005					x	8.664**
Model with 3 lags	-0.039**	0.001	0.020				x	10.205**
Model with 4 lags	-0.040**	-0.002	0.014	-0.014			x	10.558**
Model with 5 lags	-0.041**	-0.003	0.013	-0.015	-0.004		x	10.320*
Model with 6 lags	-0.038**	0.000	0.013	-0.015	-0.006	0.001	x	9.466

Panel B: Winsorizing *IMX*					
	NI_70 $_{t-1}$	*NI_70* $_{t-2}$	*NI_70* $_{t-3}$		$\chi 2$
IMX WIN 99	-0.032405**	0.006822	0.018784	x	9.924**
IMX WIN 95	-0.025086**	0.006243	0.010306	x	8.138**

Panel C: Testing the *Base Model* with different Market Sentiment Indices					
	Sentiment $_{t-1}$	*Sentiment* $_{t-2}$	*Sentiment* $_{t-3}$		$\chi 2$
IFO	0.000	0.000	0.000	x	1.238
KONBAR	0.001	0.001	-0.002753*	x	3.788
CONCLIMATE	0.026**	0.023*	-0.017	x	14.288***

Notes: This table reports results for the estimated VAR models with different variations, so as to test the robustness of Panels A-C. All regressions were run with the same set of macroeconomic control variables: The set of macroeconomic control variables includes: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates without any lag, industry turnover growth rate with a fourth lag, wages with a first and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

In addition, this paper tests three German sentiment measures regarding their relationship with the housing market. *IFO* is the business climate index for the overall economy and *CONCLIMATE* is for the construction industry, both published by the *Institute for Economic Research*. *KONBAR* refers to the business climate index published by the *DIW*. All three measures are tested in the vector autoregressive framework with the same set of control variables and lags as in the *Base Model*. As shown in Panel C in Exhibit 4.10, *IFO* and *KONBAR* do not have any significant relationship at all with the residential market returns. However, *CONCLIMATE* significantly explains the *IMX,* with the first and second lag showing statistical significance. Overall, *CONCLIMATE* Granger-causes *IMX* returns at the 1% significance level. In order to test whether the news-based sentiment measure *NI_70* contains similar information as *CONCLIMATE*, a VAR model is estimated with both measures. The two sentiment measures might explain different effects. Hence, the *Base Model* is augmented by the significant lags *CONCLIMATE(-1)* and *CONCLIMATE (-2)*.

Exhibit 4.11 confirms the robustness of the *Negative Indicator* created with the *GRESD_70*. The first lag of *NI_70* still has significant predictive power. Nonetheless, one lag of *CONCLIMATE* or two lags are included in the vector autoregressive framework. For both variations, a significant Granger Causality supports the findings. This means that the news-based sentiment measure seems to capture another form of sentiment, which is not already explained by *CONCLIMATE*. In conclusion, it is definitely worth considering sentiment measures extracted from newspaper articles by means of the dictionary-based approach, in order to improve direct German real estate market models.

Exhibit 4.11 | Robustness test with *CONCLIMATE*

	IMX			
	Model 13		Model 14	
	NI_70		*NI_70*	
IMX (-1)	0.353	***	0.380	***
	[3.86942]		[4.34213]	
IMX (-2)	-0.175	*	-0.160	*
	[-1.87650]		[-1.73492]	
IMX (-3)	-0.175	**	-0.165	**
	[-2.12899]		[-2.01639]	
Sentiment (-1)	-0.034	**	-0.036	**
	[-2.28309]		[-2.41368]	
Sentiment (-2)	0.000		-0.001	
	[0.01488]		[-0.04915]	
Sentiment (-3)	0.018		0.020	
	[1.15220]		[1.25231]	
Constant	-0.004	***	-0.004	**
	[-2.73347]		[-2.59893]	
Macroeconomic Controls	YES		YES	
CONCLIMATE (-1)	0.025	**	0.030	***
	[2.30190]		[2.93010]	
CONCLIMATE (-2)	0.013	*		
	[1.07751]			
R^2	0.606		0.602	
Adj. R^2	0.552		0.551	
Log likelihood	586.307		585.645	
Akaike AIC	-9.125		-9.130	
Schwarz SC	-8.763		-8.791	
Granger Causality				
Sentiment indicator	8.199	**	9.213	**
IMX	0.636		0.170	

Notes: This table reports results for the estimated VAR models with 3 lags, monthly IMX returns, and news-based sentiment as endogenous variables. The set of macroeconomic control variables includes: unemployment growth rate with a seventh lag, building permits and construction turnover growth rates with no lag, industry turnover growth rate with a fourth lag, wages with a first and home loan interest rate with a second lag. For brevity, the table reports the results of the real estate return equations. For brevity, the table reports the results of the real estate return equations. T-statistics are reported in square brackets below each coefficient estimate. Granger Causality values are reported for both directions. The regression is based on 129 observations from 2007M02 to 2017M10. * denotes significance at the 10% level, ** at the 5% level and *** at the 1% level.

4.7 Conclusion

In recent years, there has been increasing interest in quantifying sentiment and accordingly investigating its influence on market movements. This topic attracts ever-growing attention, due to findings consistently approving that investors' decision-making processes can be influenced by whether they feel optimistic or pessimistic about current market conditions (Bollen *et al.*, 2011). Thanks to the internet revolution, huge amounts of information are now available online. This has paved the way for a new field of sentiment analysis, namely textual analysis, which attempts to extract sentiment from various kinds of text documents. However, there have so far been few quantitative textual analysis applications in real estate. Some initial attempts have analyzed the inherent sentiment of housing news in the US and UK (Walker, 2014, 2016; Soo, 2015). Referring to Germany, research lags far behind, as no previous study has ever investigated real estate-related German text entities.

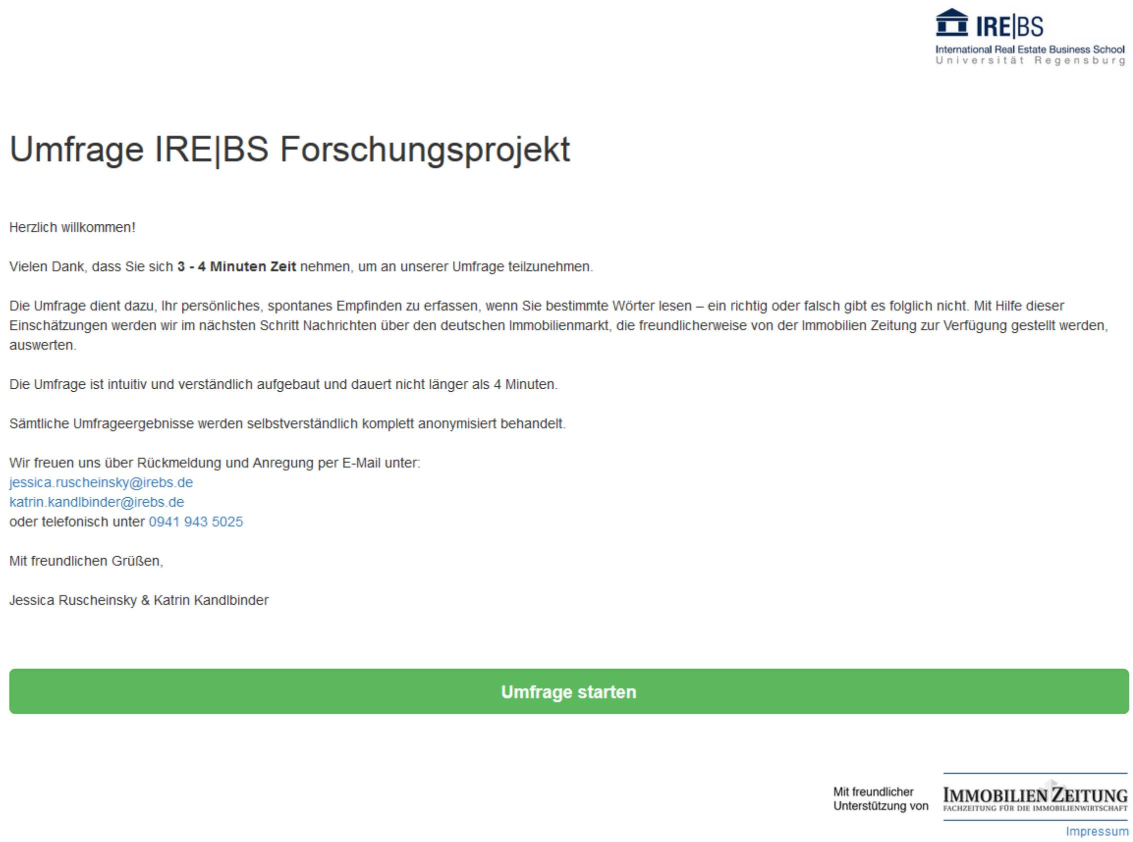
Therefore, this study was designed to investigate the relationship between news-based sentiment measures extracted by means of textual analysis and the direct residential real estate market in Germany. A fundamental prerequisite to performing the dictionary-based approach is an appropriate sentiment pre-annotated word list. As no such dictionary exists for an economic context in German, one aim of this paper was to construct the first *German Real Estate Sentiment Dictionary*. The results of the survey among about 1,700 real estate professionals resulted in a real estate-related word list with objective sentiment scores. The final *GRESD* comprises 14,137 sentiment-annotated words. Having this exceptional resource on hand, 125,462 newspaper articles published by the *Immobilien Zeitung* were analyzed with the help of the dictionary-based approach. Subsequently, the generated monthly positive and negative sentiment indicators were used to augment fundamental-based vector autoregressive models for the residential real estate market.

Most importantly, the results reveal a significant relationship between the created news-based sentiment measures and the direct real estate market. The *Negative Indicator* influences one-month-ahead *IMX* returns, even when controlling for a set of macroeconomic variables such as unemployment, building permits, construction turnover, industry turnover, wages, the home loan interest rate, and another indirect sentiment measure. However, no significant evidence can be reported for the *Positive Indicator*. This supports the notion that individuals are affected more strongly by negative rather than positive news. In order to gain deeper knowledge about the construction of suitable sentiment annotated word lists, different scopes for design were examined. It turned out that the number of words included and the sentiment intensity play a central role. Furthermore, this paper confirms that the consideration of headlines alone already generates robust sentiment measures. Comparing the domain-specific *GRESD* to general sentiment dictionaries reinforces the value and quality of the self-developed *German Real Estate Sentiment Dictionary.* Several robustness-checks confirmed the strength of the findings.

These results are not only valuable for academia but also for decision-making processes in the real estate industry. Like the forecasting results suggest, news-based sentiment measures can help anticipate future market movements. The created *German Real Estate Sentiment Dictionary* does not claim to be exhaustive. Rather, it should be seen as the groundwork for future German text-based sentiment analysis. As with Loughran and McDonald (2011), the development of a sentiment dictionary takes time and several revisions. Therefore, future work could extend the current list of words containing sentiment with regard to a real estate context. Furthermore, the *GRESD* could be applied to various text documents such as earnings press releases, annual reports, 10 Ks, analyst reports, commentaries, or IPO prospectuses and other real estate markets as well. As this approach is scalable, an even shorter aggregation frequency – weekly or even daily – could be tested if enough text entities are available.

4.8 Appendix

Appendix 4.1 | Questionnaire layout



IRE|BS International Real Estate Business School Universität Regensburg

Umfrage IRE|BS Forschungsprojekt

Herzlich willkommen!

Vielen Dank, dass Sie sich **3 - 4 Minuten Zeit** nehmen, um an unserer Umfrage teilzunehmen.

Die Umfrage dient dazu, Ihr persönliches, spontanes Empfinden zu erfassen, wenn Sie bestimmte Wörter lesen – ein richtig oder falsch gibt es folglich nicht. Mit Hilfe dieser Einschätzungen werden wir im nächsten Schritt Nachrichten über den deutschen Immobilienmarkt, die freundlicherweise von der Immobilien Zeitung zur Verfügung gestellt werden, auswerten.

Die Umfrage ist intuitiv und verständlich aufgebaut und dauert nicht länger als 4 Minuten.

Sämtliche Umfrageergebnisse werden selbstverständlich komplett anonymisiert behandelt.

Wir freuen uns über Rückmeldung und Anregung per E-Mail unter:
jessica.ruscheinsky@irebs.de
katrin.kandlbinder@irebs.de
oder telefonisch unter 0941 943 5025

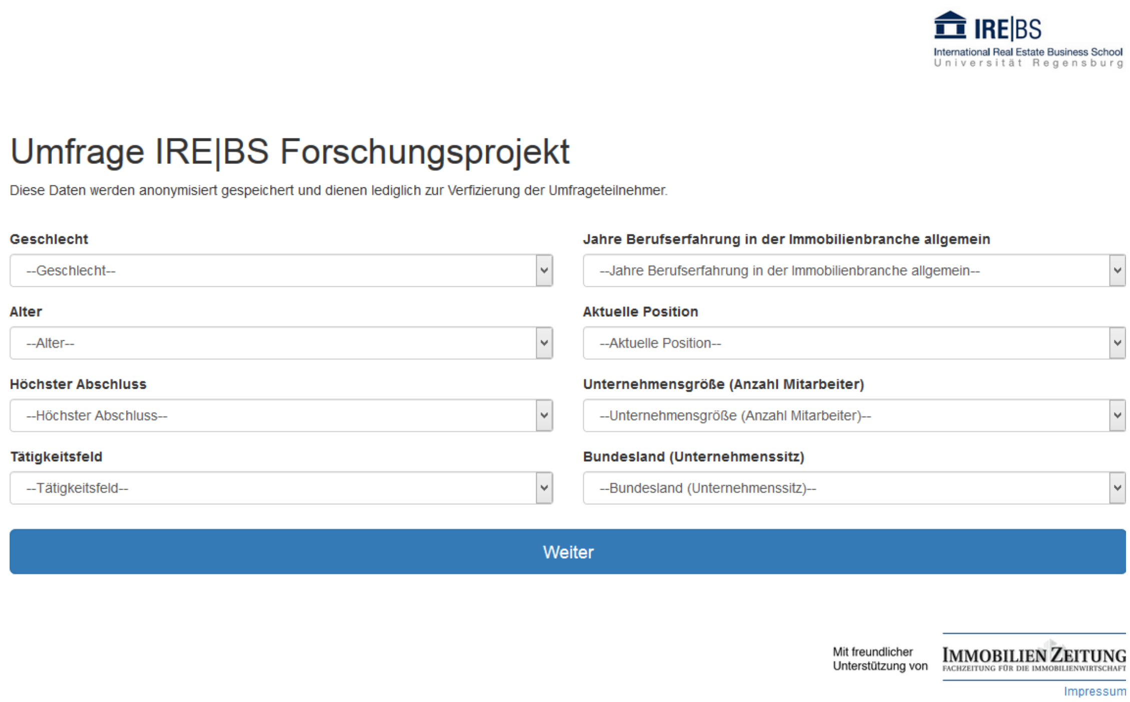
Mit freundlichen Grüßen,

Jessica Ruscheinsky & Katrin Kandlbinder

Umfrage starten

Mit freundlicher Unterstützung von IMMOBILIEN ZEITUNG FACHZEITUNG FÜR DIE IMMOBILIENWIRTSCHAFT

Impressum



IRE|BS International Real Estate Business School Universität Regensburg

Umfrage IRE|BS Forschungsprojekt

Diese Daten werden anonymisiert gespeichert und dienen lediglich zur Verfizierung der Umfrageteilnehmer.

Geschlecht
--Geschlecht--

Alter
--Alter--

Höchster Abschluss
--Höchster Abschluss--

Tätigkeitsfeld
--Tätigkeitsfeld--

Jahre Berufserfahrung in der Immobilienbranche allgemein
--Jahre Berufserfahrung in der Immobilienbranche allgemein--

Aktuelle Position
--Aktuelle Position--

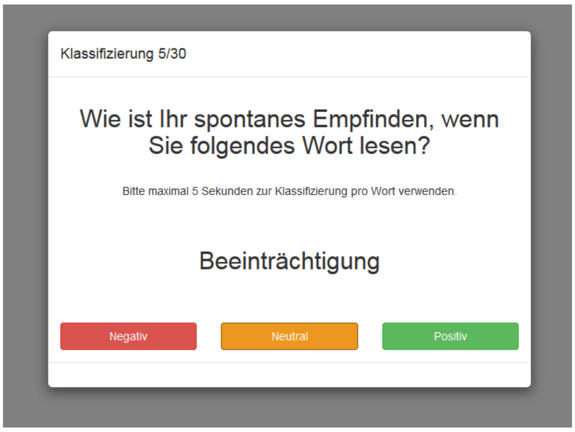
Unternehmensgröße (Anzahl Mitarbeiter)
--Unternehmensgröße (Anzahl Mitarbeiter)--

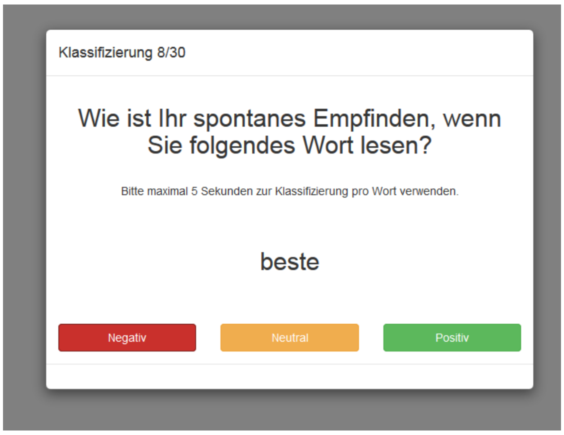
Bundesland (Unternehmenssitz)
--Bundesland (Unternehmenssitz)--

Weiter

Mit freundlicher Unterstützung von IMMOBILIEN ZEITUNG FACHZEITUNG FÜR DIE IMMOBILIENWIRTSCHAFT

Impressum


Klassifizierung 5/30
Wie ist Ihr spontanes Empfinden, wenn Sie folgendes Wort lesen?
Bitte maximal 5 Sekunden zur Klassifizierung pro Wort verwenden.
Beeinträchtigung
Negativ
Neutral
Positiv


Klassifizierung 8/30
Wie ist Ihr spontanes Empfinden, wenn Sie folgendes Wort lesen?
Bitte maximal 5 Sekunden zur Klassifizierung pro Wort verwenden.
beste
Negativ
Neutral
Positiv

Appendix 4.2 | Respondent profiles

Participants	#	#/total
female	566	34%
male	1120	66%
Total	**1686**	**100%**

Age (years)	#	#/total
20-25	132	8%
26-30	264	16%
31-35	230	14%
36-40	223	13%
41-45	208	12%
46-50	218	13%
51-55	219	13%
56-60	109	6%
61-65	46	3%
>65	37	2%
Total	**1686**	**100%**

Work experience (years)	#	#/total
0-5	485	29%
6-10	307	18%
11-15	254	15%
16-20	233	14%
> 20	407	24%
Total	**1686**	**100%**

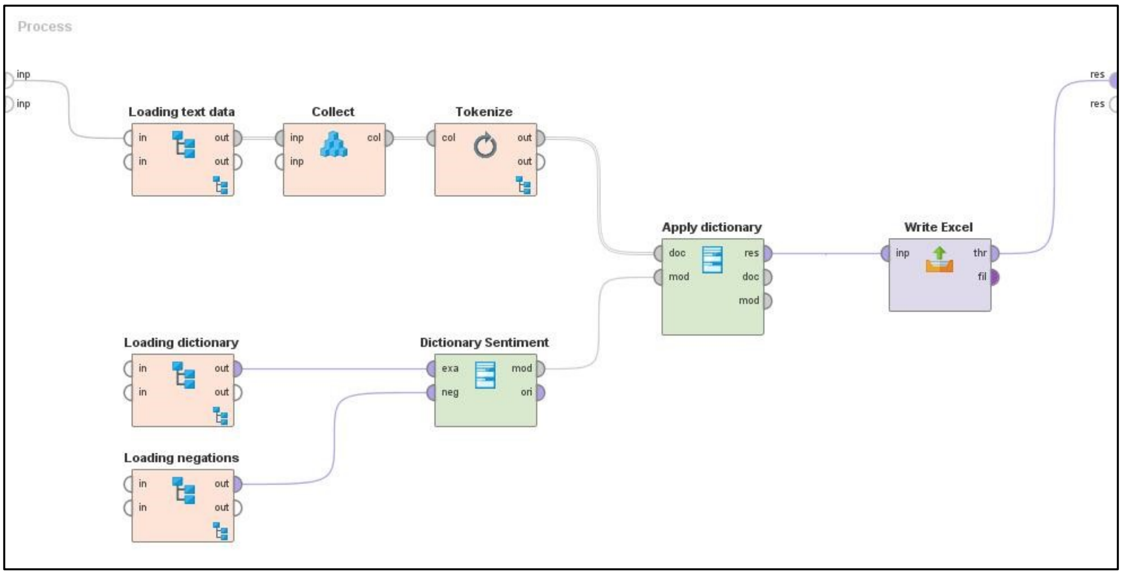
Company Size (number of employees)	#	#/total
0-5	210	12%
6-20	207	12%
21-50	180	11%
51-250	450	27%
251-500	228	14%
>500	411	24%
Total	**1686**	**100%**

Qualification	#	#/total
Diploma	557	33%
Master's degree	383	23%
Bachelor's degree	228	14%
A level (Abitur)	140	8%
Doctorate	109	6%
Advanced technical certificate (Fachhochschulreife)	79	5%
State examination	76	5%
Middle School Education (Realschule)	71	4%
Professor	34	2%
Lower Secondary Education (Hauptschulabschluss)	9	1%
Total	**1686**	**100%**

Real Estate Sector	#	#/total
Asset-, Property- & Facility Management	220	13%
Real Estate Development	181	11%
Real Estate Valuation or Consulting	173	10%
Real Estate Transactions / Acquisition	137	8%
Real Estate Finance	118	7%
Fund Management	112	7%
Real Estate Service Provider	111	7%
Real Estate Research	78	5%
Property Management	76	5%
Construction Company	66	4%
Project Management	59	3%
Human Resources	55	3%
Portfolio Management	54	3%
Architecture and Planning	49	3%
Real Estate Leasing	44	3%
Real Estate Marketing	43	3%
Real Estate Research and Teaching	39	2%
Real Estate Law	35	2%
Urban Planning	19	1%
Real Estate Tax	17	1%
Total	**1686**	**100%**

Position	#	#/total
Employee without management or operation responsibility	689	41%
Head of department	231	14%
Director	155	9%
Division Manager	146	9%
Self-employed	115	7%
Other	106	6%
Trainee / Working Student	93	6%
Partner	59	3%
Official	59	3%
Board Member	33	2%
Total	**1686**	**100%**

Appendix 4.3 | RapidMiner dictionary-based approach process



4.9 References

Antweiler, W. and Frank, M.Z. (2004), "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance, Vol. 59 No. 3, pp. 1259–1294.

Bollen, J., Mao, H. and Zeng, X.-J. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2 No. 1, pp. 1–8.

Brooks, C. and Tsolacos, S. (2010), *Real estate modelling and forecasting*, Cambridge University Press, Cambridge.

Cieleback, M. (2012), "Development of Residential Property", in Just, T. and Maennig, W. (Eds.), *Understanding German Real Estate Markets*, *Management for Professionals*, Springer, Dordrecht.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial Real Estate Valuation. Fundamentals Versus Investor Sentiment", *The Journal of Real Estate Finance and Economics*, Vol. 38 No. 1, pp. 5–37.

comScore (2016), "Media Metrix, Unique Users", *https://www.immobilienscout24.de/werbung/scout24media/plattformen/immobilienscout24.html*.

Das, S.R. and Chen, M.Y. (2007), "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web", *Management Science*, Vol. 53 No. 9, pp. 1375–1388.

Doran, J.S., Peterson, D.R. and Price, S.M. (2012), "Earnings Conference Call Content and Stock Price: The Case of REITs", *The Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 402–434.

Edelmann.ergo (2017), "Immobilien Umfrage 2016", *https://cdn.iz.de/media/documents/ansicht-immo-umfrage-2016.pdf*.

Engelberg, J. (2008), "Costly Information Processing: Evidence from Earnings Announcements", Northwestern University, *Working paper*.

Engelberg, J.E., Reed, A.V. and Ringgenberg, M.C. (2012), "How are shorts informed?", *Journal of Financial Economics*, Vol. 105 No. 2, pp. 260–278.

Ferris, S.P., Hao, Q. and Liao, M.-Y. (2013), "The Effect of Issuer Conservatism on IPO Pricing and Performance", *Review of Finance*, Vol. 17 No. 3, pp. 993–1027.

Frankel, R., Mayew, W.J. and Sun, Y. (2010), “Do pennies matter? Investor relations consequences of small negative earnings surprises”, *Review of Accounting Studies*, Vol. 15 No. 1, pp. 220–242.

Freybote, J. and Seagraves, P.A. (2017), “Heterogeneous Investor Sentiment and Institutional Real Estate Investments”, *Real Estate Economics*, Vol. 45 No. 1, pp. 154–176.

Garcia, D. (2013), “Sentiment during Recessions”, *The Journal of Finance*, Vol. 68 No. 3, pp. 1267–1300.

Heston, S.L. and Sinha, N.R. (2017), “News versus Sentiment: Predicting Stock Returns from News Stories”, *Financial Analysts Journal*, Vol. 73 No. 3, pp. 67–83.

International Monetary Fund (2017), “World Economic Outlook Database”, *https://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx*.

JLL (2016), “Global Real Estate Transparency Index 2016. Taking Real Estate Transparency to the Next Level”, *http://www.jll.com/GRETI.*

Just, T. and Maennig, W. (Eds.) (2012), *Understanding German Real Estate Markets, Management for Professionals*, Springer, Dordrecht.

Just, T., Voigtländer, M., Eisfeld, R., Henger, R.M., Hesse, M. and Toschka, A. (2017), “Wirtschaftsfaktor Immobilien 2017”.

Krosnick, J.A. and Presser, S. (2009), “Question and Questionnaire Design”, in Wright, J.D. and Marsden, P.V. (Eds.), *Handbook of Survey Research: 2nd Edition*, Elsevier, San Diego, pp. 1–81.

Li, F. (2010), “The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach”, *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049–1102.

Ling, D.C., Naranjo, A. and Scheick, B. (2014), “Investor Sentiment, Limits to Arbitrage and Private Market Returns”, *Real Estate Economics*, Vol. 42 No. 3, pp. 531–577.

Loughran, T. and McDonald, B. (2011), “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”, *The Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Loughran, T. and McDonald, B. (2015), “The Use of Word Lists in Textual Analysis”, *Journal of Behavioral Finance*, Vol. 16 No. 1, pp. 1–11.

Loughran, T. and McDonald, B. (2016), “Textual Analysis in Accounting and Finance: A Survey”, *Working paper*.

Marcato, G. and Nanda, A. (2016), “Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market”, *Journal of Real Estate Research*, Vol. 38 No. 2, pp. 165–203.

Maurer, R., Reiner, F. and Sebastian, S. (2004), “Characteristics of German Real Estate Return Distributions: Evidence from Germany and Comparison to the U.S. and U.K.”, *Journal of Real Estate Portfolio Management*, Vol. 10 No. 1, pp. 59–76.

Mayer, C.J. and Sinai, T. (2009), “U.S. House Price Dynamics and Behavioral Economics”, Policy Making insights from Behavioral Economics, *Federal Reserve Bank of Boston*.

Nakajima, M. (2011), “Understanding House-Price Dynamics”, *Business Review*, Q2, pp. 20–28.

Newswhip (2017), “How Long Are The Most Shared Stories On Social Media?”, *https://www.newswhip.com/2017/01/long-shared-stories-social-media/.*

Price, S.M., Doran, J.S., Peterson, D.R. and Bliss, B.A. (2012), “Earnings conference calls and stock returns: The incremental informativeness of textual tone”, *Journal of Banking & Finance*, Vol. 36 No. 4, pp. 992–1011.

Remus, R., Quasthoff, U. and Heyer, G. (2010), “SentiWS – a Publicly Available German-language Resource for Sentiment Analysis”, *LREC*, pp. 1168–1171.

Rozin, P. and Royzman, E.B. (2001), “Negativity Bias, Negativity Dominance, and Contagion”, *Personality and Social Psychology Review*, Vol. 5 No. 4, pp. 296–320.

Ruscheinsky, J.R., Lang, M. and Schäfers, W. (2018), “Real Estate Media Sentiment Through Textual Analysis”, *Journal of Property Investment & Finance*, Vol. 36 No. 5.

Schulte, K., Rottke, N. and Pitschke, C. (2005), “Transparency in the German real estate market”, *Journal of Property Investment & Finance*, Vol. 23 No. 1, pp. 90–108.

Schulte, K.-W., Gier, S., Evans, A. and Lee, A. (2011), *Wörterbuch Immobilienwirtschaft, Immobilienzeitung-Edition,* 4.th ed., IZ Immobilienzeitung, Wiesbaden.

Soo, C.K. (2015), “Quantifying Animal Spirits: News Media and Sentiment in the Housing Market”, *Working paper*.

Strapparava, C. and Mihalcea, R. (2008), "Learning to identify emotions in text", in Wainwright, R.L. (Ed.), *Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil*, ACM, New York, NY, pp. 1556-1560.

Tashman, L.J. (2000), "Out-of-sample tests of forecasting accuracy: an analysis and review", *International Journal of Forecasting*, Vol. 16 No. 4, pp. 437–450.

Taylor, J.B. (2013), *Getting off track: How government actions and interventions caused, prolonged, and worsened the financial crisis*, Hoover Press.

Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008), "More Than Words: Quantifying Language to Measure Firm's Fundamentals", *The Journal of Finance*, Vol. 63 No. 3, pp. 1437–1467.

Theil, H. (1966), *Applied Economic Forecasting*, Noth-Holland, Amsterdam.

Theil, H. (1971), *Principles of Econometrics*, Wiley, New York, NY.

Walker, C.B. (2014), "Housing booms and media coverage", *Applied Economics*, Vol. 46 No. 32, pp. 3954–3967.

Walker, C.B. (2016), "The direction of media influence: Real-estate numes and the stock market", *Journal of Behavioral and Experimental Finance*, Vol. 10, pp. 20–31.

Waltinger, U. (2010), "GermanPolarityClues - a Lexical Resource for German Sentiment Analysis", *LREC*, pp. 1638–1642.

Winson, K. (2017), "Sentiments and Semantics: A review of the content analysis literature in the era of big data", *Journal of Real Estate Literature,* Manuscript Draft.

5 Conclusion

In the course of this dissertation, the demand for information, information availability, and the supply of information in both direct and indirect real estate markets have been analyzed. By applying different empirical methods in each paper, the thesis demonstrates the usefulness of emerging information sources from the internet and the increasing digitalization.

The following section provides a summary of the three papers comprising the motivation, research design and main findings of each contributing paper. The dissertation closes with some final remarks as well as with suggestions for potential further research.

5.1 Executive Summary

Paper 1 | Intraday Online Information Demand and its Relationship with REIT Prices

The first paper of this dissertation investigates the intraday relationship between Google search volumes and the indirect real estate market proxied by US REITs. Using search engines like Google has become synonymous with accessing the internet – the most important source for acquiring information nowadays. Given that investors seek information before making investment decisions, search volumes provided by Google Trends have the ability to reveal market participants' interests.

Hypothetical trading strategies for the Top 20 MSCI US REITs are developed, based on hourly changes in Google search volume for 78 trading days. In order to find out whether Google search volumes have the ability to successfully predict future REIT price movements, the performance of search-based trading strategies are compared to the results of a buy-and-hold strategy. Furthermore, this study investigates in which market circumstances (falling, rising, or stagnant markets) and before which trading signal (buy or sell) the Google trading strategy has the best forecasting abilities. Moreover, the issue of correlation and causality between search volumes and REIT prices is examined. This is the first paper to use intraday search volumes and prices in order to overcome the measurement imprecisions of weekly data and to gain more granular insights into information search behavior.

The most striking result is that the formulated Google trading strategies produce better performance measures than the tested benchmarks. The Google trading strategy outperforms the benchmark by 7.37 percentage points on average. The best performing search-volume-based strategy would have gained a remarkable outperformance of 45.60 percentage points, compared to the performance of the simple buy-and-hold strategy. Additionally, it has been shown that Google trading strategies perform considerably better during declining market phases. As shown by the hit rate for long and short signals, market participants tend to search more intensively for company-related search terms before buying a stock, rather than before selling it. At an aggregate level, the results show a statistically negative correlation of -0.11 and a causality flow from prices to search volumes, but individually, these findings cannot be confirmed.

In conclusion, this research shows that hourly Google Trends search volumes represent a successful new way to predict short-term REIT prices. This can provide researchers with profound insights into how information is acquired and how it is used to make investment decisions in REIT markets.

Paper 2 | Leveling the Playing Field: Out-of-Town Buyer Premiums in US Housing Markets Over Time

The second paper is the first to analyze the out-of-town buyer premium with respect to improvements in information availability over a ten year window. The existing literature has found that distant buyers pay higher prices for real estate than local buyers. The reasons are attributed to the fact that out-of-town buyers tend to face higher search costs and are therefore informationally disadvantaged. Furthermore, they might have upwardly biased expectations of property values. The aim of this paper is to investigate whether the enormous increase in information availability over the last decade helped to equalize the information level between heterogeneous buyers.

By incorporating both a theoretical search model and a hedonic regression model for 2005 and 2015, this study explores whether out-of-town buyers do in fact pay a premium and why, and whether this premium has decreased due to better information availability through the internet. In decomposing what causes the premium, this article investigates three potential premium sources, namely search costs/information asymmetries (distance), biased beliefs (anchoring), and different income levels (wealth).

The results indicate that distant buyers continue to pay higher prices for real estate than local buyers in 2015, although the premium in 2015 was lower than in 2005. This decline may be due to better information availability. The premium is caused mainly by search costs (distance), whereas behavioral bias in the form of anchoring tends to play a less important role, as the coefficient is statistically significant, but very small. The premium caused by the average personal income (wealth) of the buyer's origin is only statistically significant in 2005. Propensity score matching regarding the property characteristics in 2005 and 2015 and various robustness checks validate these findings. The intercounty comparison with San Diego and San Francisco County reveals that further work is needed in order to better isolate the causes of the out-of-town buyer premium in real estate markets.

Overall, this study confirms that the internet presumably does contribute to equalizing prices paid by distant buyers, as the premium decreased from 2005 to 2015. The enormous increase in information availability due to the internet is shown to cause changes in prices, which will ultimately affect economic outcomes as well.

Paper 3 | Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany

The third paper of this dissertation is the first application of dictionary-based textual analysis in the German real estate market. International research has confirmed that sentiment is valuable for explaining real estate market movements, and that textual analysis is an appropriate technique for extracting and quantifying sentiment. However, this approach has not yet been applied to the German real estate market.

In order to apply textual analysis, namely the dictionary-based approach, the *German Real Estate Sentiment Dictionary* was developed. By conducting a large scale online survey among approximately 1,700 real estate professionals, presumably sentiment-loaded words were objectively classified as positive or negative. This discipline-specific sentiment dictionary enables examining 125,462 newspaper articles from 2007 to 2017 published in the *Immobilien Zeitung*. A vector autoregressive framework and out-of-sample forecasts are utilized to analyze the dynamic relationship between the self-created sentiment measures and the German housing market, which is proxied by the *IMX* index.

Most importantly, the results reveal strong and robust evidence of predictive power of the news-based sentiment measures regarding future housing price movements. Even when controlling for various macroeconomic variables, the negative sentiment indicator significantly influences and Granger-causes future *IMX* returns. For the positive sentiment indicator, no significant influence can be reported. This finding confirms the notion that market participants are affected more strongly by negative rather than positive news. Additionally, it is shown that even analyzing only the headlines of each newspaper articles yields sufficiently powerful sentiment measures and statistically significant housing price predictions. Comparing sentiment-augmented vs. non-sentiment VAR Models, the results of the out-of-sample forecasting indicate that sentiment measures indeed enhance forecasting accuracy.

In conclusion, this paper implies that textual analysis can be applied successfully to the German real estate market. It provides valuable insights which enable a better understanding of the influences on German residential market movements, that are not based solely on fundamental changes. Most notably, the objectively validated *German Real Estate Sentiment Dictionary* lays the foundation for future textual analysis applications in the German real estate market.

5.2 Final Remarks and Future Research

The rapidly growing provision of information is likely to continue for several reasons, including the increasing number of social media users, decreasing data-storage costs, and highly sophisticated data processing technologies. The information revolution has evidently already changed our social lives and the ways in which we procure and disseminate information. Economies will become even more integrated, due to technological innovation and the reduction of barriers across different market segments. Naturally, the question arises of whether these developments are perceived as fascinating or frightening. On the one hand, the actions in our everyday lives will become even more traceable, measurable, and will be analyzed thoroughly. Hence, the sheer amount of personal data inevitably creates more entry points for hackers and renders sensitive information vulnerable. On the other hand, we will be able to make better informed, more accurate, and more timely decisions, which will improve our lives.

In summary, the three papers within this dissertation investigate three different informational phenomena and their effects on real estate markets. In the first paper, information demand measured by the relatively new internet source Google Trends is applied to reveal sentiment in indirect real estate markets. Traditionally, researchers proxy investor sentiment with market-based measures such as abnormal trading volumes, IPO first day returns or implied volatilities, or they use survey-based indices (Da *et al.*, 2014). Nowadays, driven by the internet revolution, the innovative tool Google Trends enables to directly measure investor attention and has turned out to be a good sentiment indicator. In the second paper, changes in information availability were examined by comparing premiums paid by out-of-town buyers over a ten year window. One decisive aspect of the information revolution is that information can be acquired at lower cost. Thus, with decreasing search costs and increasing information availability, the prices for comparable properties should be more equal for all kinds of buyers (Clauretie and Thistle, 2007). The third paper concentrates on information supply, by using newspaper articles together with the dictionary-based approach to identify sentiment levels. The growing availability of digitized textual sources and the constant production of new digital-born sources in the course of the information revolution, create new application possibilities for textual analysis. As textual analysis is an appropriate technique for extracting sentiment from textual sources, it has attracted much attention from academia (Tetlock, 2007).

However, some promising research questions remain to be answered. Regarding the first paper, technological improvements in the tool Google Trends have made it possible to examine hourly internet search behavior instead of weekly search-volume frequencies. This increased temporal granularity is especially helpful, when analyzing stocks or REITs. For the direct real estate markets, increasing geographical granularity from country to state or even city-level would be a potential next step. This improvement should yield valuable insights into both residential and commercial real estate markets, as they are known to be geographically segmented. Nevertheless, the availability of data on fundamentals, which also limited the scope of the first study, will remain a challenge for further research.

Regarding the second paper, there remains the question of how information intermediaries in the form of real estate agents, influence pricing outcomes. Especially in housing markets, real estate agents play a crucial role, as many buyers still rely on their support to make one of the most important investment decisions in their lives (Ling *et al.*, 2016). Other factors like condominium fees for amenities or the number of international companies which attract high potentials or expats and hereby disregard any pricing levels may help to explain different pricing outcomes in different counties. Furthermore, a deeper analysis of spatial distance as a proxy for information levels might be of considerable interest. The distance between REIT headquarters and the properties they acquire, might affect the REIT's long-term performance, due to informational advantages when buying and managing the properties.

With respect to the third paper, the application of newly available data and techniques to the German real estate markets probably represents the area with the greatest need and potential for further research. Until now, the German real estate market has continued to lag behind other markets like the US or the UK in terms of data availability and accessibility. However, the JLL market transparency index (JLL, 2016) has ranked the German real estate market as "highly transparent" for the first time, due to its growth in the listed sector. The monthly sentiment indices could only be applied to a housing-offer price index so far. An application to a monthly transaction-based price index both for residential and commercial real estate market would take German real estate research to the next level. Regarding the indirect real estate market, a valuable expansion would be the application of text mining methods to annual reports of German listed real estate companies in order to predict their performance. Furthermore, other textual analysis techniques like machine learning have not yet been tested on German real estate newspaper articles with the aim of detecting sentiment.

In summary, the three contributing papers of this dissertation reveal that newly emerging information sources and channels, combined with innovative analyzing tools, enable more accurate predictions of future market movements, both in indirect and direct real estate markets. Additionally, the increase in information availability due to the internet contributes to equalizing the information level between heterogeneous market participants. Hence, this dissertation provides evidence that information is still one of the most valuable resources, even if the procurement and processing has changed considerably. To make a long story short:

Information remains the key.

5.3 References

Clauretie, T.M. and Thistle, P.D. (2007), "The Effect of Time-on-Market and Location on Search Costs and Anchoring. The Case of Single-Family Properties", *The Journal of Real Estate Finance and Economics*, Vol. 35 No. 2, pp. 181–196.

Da, Z., Engelberg, J. and Gao, P. (2014), "The Sum of All FEARS Investor Sentiment and Asset Prices", *Review of Financial Studies*, Vol. 28 No. 1, pp. 1–32.

JLL (2016), "Global Real Estate Transparency Index 2016. Taking Real Estate Transparency to the Next Level", *http://www.jll.com/GRETI*.

Ling, D.C., Naranjo, A. and Petrova, M.T. (2016), "Search Costs, Behavioral Biases, and Information Intermediary Effects", *The Journal of Real Estate Finance and Economics*, Vol. 35 No. 9, pp. 2330–2365.

Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.